IIT Madras

ONLINE DEGREE

Statistics for Data Science -1
Introduction and types of data

Usha Mohan

Indian Institute of Technology Madras

Learning objectives

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
   - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
   - ▶ Understand cross-sectional versus time-series data.
   - ▶ Measurement scales
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.

What is Statistics?

Definition

Statistics[1] is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

---

[1]Ross, Sheldon M. Introductory statistics. Academic Press, 2017.

Major branches of statistics

1. Description

Definition

The part of statistics concerned with the description and summarization of data is called descriptive statistics.

2. Inference

Definition

The part of statistics concerned with the drawing of conclusions from data is called inferential statistics.

▶ To be able to draw a conclusion from the data, we must take into account the possibility of chance- introduction to probability.

Population and sample

Suppose we are interested in knowing

▶ The percentage of all students in India who have passed their Class 12 exams and study engineering.

▶ The prices of all houses in Tamil Nadu.

▶ The total sales of all cars in India in the year 2019.

▶ The age distribution of people who visit a city Mall in a particular month.

Population and sample

Definition
The total collection of all the elements that we are interested in is called a population.

Definition
A subgroup of the population that will be studied in detail is called a sample.

Purpose of statistical analysis

▶ If the purpose of the analysis[2] is to examine and explore information for its own intrinsic interest only, the study is descriptive.

▶ If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the study is inferential.

▶ A descriptive study may be performed either on a sample or on a population.

▶ When an inference is made about the population, based on information obtained from the sample, does the study become inferential.

---

[2]Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Summary

- ▶ Descriptive statistics
- ▶ Inferential statistics
- ▶ Population and sample

Learning objectives

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
   - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
   - ▶ Understand cross-sectional versus time-series data.
   - ▶ Measurement scales
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.

What is Data

In order to learn something, we need to collect data.

Definition

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

▶ Statistics relies on data, information that is around us.

Why do we collect Data

▶ Interested in the characteristics of some group or groups of people, places, things, or events.

▶ Example: To know about temperatures in a particular month in Chennai, India.

▶ Example: To know about the marks obtained by students in their Class 12.

▶ To know how many people like a new song/product/video-collected through comments.

Data collection

- Data available: published data.
- Data not available: need to collect, generate data.

We assume data is available and our objective is to do a statistical analysis of available data.

Unstructured and structured data

▶ For the information in a database to be useful, we must know the context of the numbers and text it holds.

▶ When they are scattered about with no structure, the information is of very little use.

▶ Hence, we need to organize data

Dataset

- ▶ A structured collection of data.
- ▶ it is a collection of values-could be numbers, names, roll numbers.
- ▶ https://docs.google.com/spreadsheets/d/
  15nJvZ-xBZDGb0oii-NCvSIY4fETotXcJdm5pV1Fq2aI/
  edit?usp=sharing
- ▶ https://docs.google.com/spreadsheets/d/
  1qZWmXsIpFx10srpFcmj9DPA961UMbTXkCiUr_SxBYq4/
  edit?usp=sharing
- ▶ https://docs.google.com/spreadsheets/d/
  1lrmhe-E0A2LWpTB9cBK9dm-sL2SPVXYZl0MJHI6vqhM/
  edit?usp=sharing

Variables and cases

- ▶ Case (observation): A unit from which data are collected
- ▶ Variable:
  - ▶ Intuitive: A variable is that "varies".
  - ▶ Formally: A characteristic or attribute that varies across all units.
- ▶ In our school data set:
  - ▶ Case: each student
  - ▶ Variable: Name, marks obtained, Board etc.
- ▶ Rows represent cases: for each case, same attribute is recorded
- ▶ Columns represent variables: For each variables, same type of value for each case is recorded.

Summary

We have organized data in a spreadsheet into a table

Each variable must have its own column.
Each observation must have its own row.

Learning objectives

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
   - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
   - ▶ Understand cross-sectional versus time-series data.
   - ▶ Measurement scales
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.

Statistics for Data Science -1
└ Classification of data
  └ Categorical and numerical

Categorical and numerical
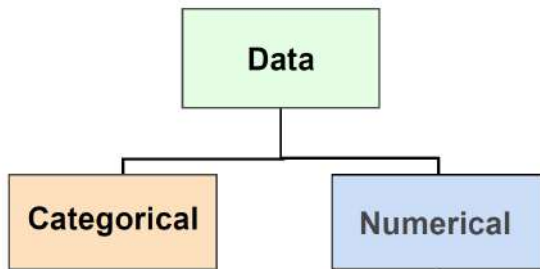
Categorical and numerical variables

- ▶ Categorical data
  - ▶ Also called qualitative variables.
  - ▶ Identify group membership
- ▶ Numerical data
  - ▶ Also called quantitative variables.
  - ▶ Describe numerical properties of cases
  - ▶ Have measurement units
- ▶ Measurement units: Scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in centimeters, etc.
  - ▶ The data that make up a numerical variable in a data table must share a common unit.

Cross-sectional and time-series data

- ▶ Time series - data recorded over time
- ▶ Timeplot – graph of a time series showing values in chronological order
- ▶ Cross-sectional - data observed at the same time

Statistics for Data Science -1
└─ Classification of data
  └─ Cross-sectional versus time-series data

Time-series data- Example

Date	Potato		
	Qty(KG)	cost (Rs.)	Selling price(Rs.)
01-Mar	0	21	24
02-Mar	1350	20.05	24
03-Mar	675	20.5	24
04-Mar	0	NA	NA
05-Mar	675	20.8	24
06-Mar	675	21.25	24
08-Mar	20	20.5	24
09-Mar	900	20.5	24
10-Mar	900	20.5	24
11-Mar	0	NA	NA
12-Mar	900	20.3	24
13-Mar	1125	19.4	22
15-Mar	1125	18.8	22
16-Mar	1125	19.4	22
17-Mar	1125	19.25	22
18-Mar	1125	20.3	24
19-Mar	1125	19.8	24
20-Mar	675	21.25	24
22-Mar	675	20.5	24
23-Mar	0	NA	NA
24-Mar	0	NA	NA
25-Mar	675	19.6	24
26-Mar	675	19.7	24
27-Mar	1125	19.3	24
29-Mar	540	20.6	26
30-Mar	0		28

Statistics for Data Science -1
└─ Classification of data
   └─ Cross-sectional versus time-series data

Summary

▶ Classify data as categorical or numerical.

▶ For numerical data, find out unit of measurement.

▶ Check whether data is collected at a point of time (cross-sectional data) or over time (time-series data).

Learning objectives

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
   - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
   - ▶ Understand cross-sectional versus time-series data.
   - ▶ Measurement scales
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.

Statistics for Data Science -1
└─ Classification of data
  └─ Scales of measurement

Scales of measurement

- ▶ Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio.

Statistics for Data Science -1
└─ Classification of data
　└─ Scales of measurement

Nominal scale of measurement

- ▶ When the data for a variable consist of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a nominal scale.
- ▶ Examples: Name, Board, Gender, Blood group etc.
- ▶ Sometimes nominal variables might be numerically coded.
  - ▶ For example: We might code Men as 1 and Women as 2. Or Code Men as 3 and Women as 1. Both codes are valid.
- ▶ There is no ordering in the variable.
- ▶ | Nominal: name categories without implying order |

Ordinal scale of measurement

▶ Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered a ordinal scale.

▶ Each customer who visits a restaurant provides a service rating of excellent, good, or poor.

  ▶ The data obtained are the labels—excellent, good, or poor—the data have the properties of nominal data.

  ▶ In addition, the data can be ranked, or ordered, with respect to the service quality.

▶ Ordinal – name categories that can be ordered

Interval scale of measurement

- ▶ If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is interval scale.

- ▶ Interval data are always numeric. Can find out difference between any two values.

- ▶ Ratios of values have no meaning here because the value of zero is arbitrary.

- ▶ Interval:

  numerical values that can be added/subtracted (no absolute zero)

Statistics for Data Science -1
└─ Classification of data
   └─ Scales of measurement

Example: temperature

- ▶ Suppose the response to a question on how hot the day is comfortable and uncomfortable, then the temperature as a variable is nominal.
- ▶ Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is ordinal.
- ▶ Example: Consider a AC room where temperature is set at 20°C and the temperature outside the room is 40°C. It is correct to say that the difference in temperature is 20°C, but it is incorrect to say that the outdoors is twice as hot as indoors.
- ▶ Temperature in degrees Fahrenheit or degrees centigrade is an interval variable. No absolute zero.

	Celsius	Fahrenheit
Freezing point	0	32
Boiling point	100	212

Statistics for Data Science -1
└─ Classification of data
   └─ Scales of measurement

Summary

1. Why are jersey number, mobile number, and pincode nominal scales of measurement?
Answer: A categorical variable is said to have a nominal scale of measurement if there is no order in the categories. Generally, names and labels do not have an order. Hence, they are said to have nominal scale of measurement.

For instance, consider the jersey numbers of cricket players. There is no concept of ordering for these jersey numbers. As you can see, Dhoni's jersey number 7 is in no way greater than or lesser than Gambhir's jersey number 5. Further, it is meaningless to perform mathematical operations on any two jersey numbers. A player with jersey number 12 is not the sum of two players whose jersey numbers add up to 12 (say jersey numbers 5 and 7).

Similarly, even though mobile numbers and pin codes are numbers, it is meaningless to associate an order to these numbers. There is no order in pin codes 100002, 500001, 500002 i.e., pin code 100002 is neither greater nor lesser than 500001 or 500002. However, if the numbers 100002, 500001 and 500002 represent money in rupees, then the ordering $100002 < 500001 < 500002$ makes sense. In this case, we have an order. The same explanation applies to mobile numbers too.

2. How does the conversion of interval scale of measurement to ratio scale of measurement by subtraction work?
Answer: The conversion of interval scale to ratio scale by a subtraction operation is an application of what Prof. Usha Mohan has taught in her lectures. Variables with an interval scale of measurement have a fixed unit of measure and do not have an absolute zero (as in the case of degree Celsius). So, multiplication and division operations can not be performed on these variables.
Variables with a ratio scale of measurement have a fixed unit of measure and absolute zero (temperature in Kelvin). So division and multiplication operations can be performed.

The only difference between ratio and interval scales is with respect to absolute zero. If a variable with an interval scale of measurement has absolute zero, it will have a ratio scale of measurement. Absolute zero exists when we subtract two numerical values with an interval scale of measurement. For example, in Restaurant 1 and Restaurant 2, the rating given by users is recorded. The rating given by a user should be an integer from 1 to 5. Since the ratings given by users have a fixed measure and no absolute zero, it follows that the rating is an interval scale.
However, if we are interested in the difference (absolute value) between ratings given by the same user, then the new variable can take values from 0 to 4. This variable has

absolute zero. So, it has a ratio scale of measurement. Table 1 gives an example of ratings.

S.No.	User	Restaurant 1	Restaurant 2	Difference in rating
1.	User 1	4	1	3
2.	User 2	3	5	2
3.	User 3	5	5	0
4.	User 4	3	4	1
5.	User 5	5	3	2

Table 1: Restaurant ratings

3. Why is it that the interval scale of measurement can take negative values but not the ratio scale of measurement?
Answer: Interval scales do not have absolute zero and can take negative values, whereas ratio scales do not have negative values.
In the Kelvin scale, temperature cannot be less than zero and hence cannot be negative. In other words, zero Kelvin is the minimum value of temperature possible and hence temperature in Kelvin scale has ratio scale of measurement.
In the case of temperature measured in Celsius scale, the values can be negative. It can take up values as low as -273 degree Celsius.

4. Give an example of interval scale other than the temperature example.
Answer: Any numerical variable will have an interval scale of measurement if it has all the properties of ordinal scale and the difference between the values is expressed as a fixed unit of measure.
Consider the example of GPA (Grade Point Average) received by students based on their performance in academics. The GPA is in the range [4,10]. We see that although there is ordering in GPA, it doesn't have an absolute zero. We can not say that 8.0 GPA is twice as good as 4.0 GPA. So, Grade Point Average (GPA) has an interval scale of measurement.

5. What is the difference between ordinal scale of measurement and nominal scale of measurement?
Answer: For a categorical variable to have an ordinal scale, irrespective of frequencies there should be order (rank) in the categories.
The feedback given by customers in a fruit shop is either *Good*, *Bad* or *Average*. On a Sunday, the ratings given by a set of 45 customers are as follows:

- Good - By 30 customers
- Bad - By 15 customers
- Average - By 5 customers

The descending order based on frequencies is Good > Bad > Average. But we know that the original order of categories is Good > Average > Bad. The shop owner considers Good as highest and Bad as lowest, with Average being in between Good and Bad. Hence, ratings by the customer is an ordinal scale of measurement. We now give an example of nominal scale of measurement. Table 2 gives the quantity of each fruit sold in a fruit shop on a particular day.

S.No.	Item sold	Quantity (kg)
1.	Kiwi	5
2.	Apple	14
3.	Orange	20

Table 2: Fruit shop data

Even though more oranges were sold than either apples or kiwis, we know that orange is neither greater than nor smaller than apple or kiwi. The shop owner considers every fruit the same i.e., there is no order among the fruits. Hence, item sold is a nominal scale of measurement.

6. What is the difference between inferential statistics and descriptive statistics?
Answer: Descriptive statistics deals with describing and summarizing the given data using certain parameters while inferential statistics deals with arriving at conclusions from the data. For example, if you want to describe the performance of students in an exam conducted yesterday, then it comes under descriptive statistics. On the other hand, if you want to infer the performance of the students in the exam to be conducted tomorrow, using the data of the exam conducted yesterday, then it comes under inferential statistics.

7. What is the difference between cross-sectional and time series data?
Answer: The temperature in all major cities of India recorded at a particular instant of time is cross-sectional data. On the other hand, the temperature in Delhi recorded throughout the day is time series data. That is, if the values in the dataset vary with respect to space, with the time being constant, it is cross-sectional data. If the data varies with respect to time for a particular entity in space, it is time series data.

8. What is the difference between ordinal scale of measurement and interval scale of measurement?
Answer: In the case of ordinal scale of measurement, the difference between two consecutive pairs of values need not be the same whereas in the case of interval scale of measurement, the difference between two consecutive values is a fixed unit of measure. In the example given in Table 1, the possible ratings are integers from 1 to 5. The ordered values 1, 2, 3, 4, 5 have a difference of 1 between any two consecutive values. Therefore, the ratings dataset has an interval scale of measurement.
If the options for ratings are Good, Average, and Bad, then it has an ordinal scale

of measurement since we do not know for sure that the difference between Good and Average is same as the difference between Average and Bad.

Review

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
   - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
   - ▶ Understand cross-sectional versus time-series data.
   - ▶ Measurement scales

Frequency distributions
  Relative frequency distributions

Frequency distributions

Definition

A *frequency distribution*[1] of qualitative data is a listing of the distinct values and their frequencies.

Each row of a frequency table lists a category along with the number of cases in this category.

---

[1]Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Example

Construct a frequency table for the given data

1. A,A,B,C,A,D,A,B,D,C
2. A,A,B,C,A,D,A,B,D,C,A,B,C,D,A
3. A,A,B,C,A,A,B,B,D,C,A,B,C,D,B
4. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A,C,D,D

Construct a frequency distribution

The steps to construct a frequency distribution[2]

Step 1 List the distinct values of the observations in the data set in the first column of a table.

Step 2 For each observation, place a tally mark in the second column of the table in the row of the appropriate distinct value.

Step 3 Count the tallies for each distinct value and record the totals in the third column of the table.

[2]Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Example

1. A,A,B,C,A,D,A,B,D,C

Category	Tally mark	Frequency
A		
B		
C		
D		
Total		

2. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A

Category	Tally mark	Frequency
A		
B		
C		
D		
Total		

Example

3. A,B,B,C,A,D,B,B,D,C, A,B,C,D,B

Category	Tally mark	Frequency
A		
B		
C		
D		
Total		

4. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A,C,D,D

Category	Tally mark	Frequency
A		
B		
C		
D		
Total		

Frequency table in a googlesheet

Step 1 Select/Highlight the cells having data you want to visualize.

Step 2 In the Formatting bar click on the Data option.

Step 3 In the Data option go to Pivot Table option and create a new sheet.

Step 4 After creating Pivot Table go in Pivot Table Editor and in that first add rows and then values.

Relative frequency

Definition
The ratio of the frequency to the total number of observations is called relative frequency

▶ The steps to construct a relative frequency distribution

Step 1 Obtain a frequency distribution of the data.

Step 2 Divide each frequency by the total number of observations.

Example

1. A,A,B,C,A,D,A,B,D,C

Category	Tally mark	Frequency	Relative frequency
A	ⅠⅠⅠⅠ	4	
B	ⅠⅠ	2	
C	ⅠⅠ	2	
D	ⅠⅠ	2	
Total		10	

2. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A

Category	Tally mark	Frequency	Relative frequency
A	Ⅲ|Ⅰ	6	
B	ⅠⅠⅠ	3	
C	ⅠⅠⅠ	3	
D	ⅠⅠⅠ	3	
Total		15	

Why relative frequency?

- ▶ For comparing two data sets.
- ▶ Because relative frequencies always fall between 0 and 1, they provide a standard for comparison.

Example

3. A,B,B,C,A,D,B,B,D,C, A,B,C,D,B

Category	Tally mark	Frequency	Relative frequency
A	|||	3	
B	₩|	6	
C	|||	3	
D	|||	3	
Total		15	

4. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A,C,D,D

Category	Tally mark	Frequency	Relative frequency
A	₩|	6	
B	|||	3	
C	||||	4	
D	₩	5	
Total		18	

Sectional summary

1. Constructing a frequency table.
2. Notion of relative frequency and constructing a relative frequency table.

Charts of categorical data

▶ The two most common displays of a categorical variable are a bar chart and a pie chart.

▶ Both describe a categorical variable by displaying its frequency table.

Pie charts

Definition
A pie chart is a circle divided into pieces proportional to the relative frequencies of the qualitative data.

▶ The steps to construct a pie-chart[1]

Step 1 Obtain a relative-frequency distribution of the data.

Step 2 Divide a circle into pieces proportional to the relative frequencies.

Step 3 Label the slices with the distinct values and their relative frequencies.

---

[1]Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Example

Use a protractor and the fact that there are $360°$ in a circle. Thus, for example, the first slice of the circle is obtained by marking off $0.4 \times 360 = 144°$.

1. A,A,B,C,A,D,A,B,D,C

Category	Tally mark	Freq	Rel freq	Degrees
A	||||	4	0.4	144
B	||	2	0.2	72
C	||	2	0.2	72
D	||	2	0.2	72
Total		10	1	$360°$

Example

1. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A

Category	Tally mark	Freq	Relative freq	Degrees
A	⦀⦀	6	0.4	144
B	⦀⦀	3	0.2	72
C	⦀⦀	3	0.2	72
D	⦀⦀	3	0.2	72
Total		15	1	360°

Pie chart in a google sheet

Step 1 Select/Highlight the cells having data you want to visualize.

Step 2 Click the Insert Chart option in Google Sheets toolbar.

Step 3 Change the visualization type in Chart editor.

Step 4 Select in Chart Editor Chart type to Pie chart.

Sectional summary

1. A pie chart is used to show the proportions of a categorical variable.
2. A pie chart is a good way to show that one category makes up more than half of the total.

Bar chart

Definition

A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (or frequencies or percents) of those values on a vertical axis. The frequency/relative frequency of each distinct value is represented by a vertical bar whose height is equal to the frequency/relative frequency of that value. The bars should be positioned so that they do not touch each other.

Steps to construct a bar chart

To Construct a Bar Chart[2]

Step 1 Obtain a frequency/relative-frequency distribution of the data.

Step 2 Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies/relative frequencies.

Step 3 For each distinct value, construct a vertical bar whose height equals the frequency/relative frequency of that value.

Step 4 Label the bars with the distinct values, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" /"Relative frequency."

---

[2]Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Bar chart in a google sheet

Step 1 Select/Highlight the cells having data you want to visualize.

Step 2 Click the Insert Chart option in Google Sheets toolbar.

Step 3 Change the visualization type in Chart editor.

Step 4 Select in Chart Editor Chart type to Bar chart.

Pareto charts

Definition

When the categories in a bar chart are sorted by frequency, the bar chart is sometimes called a Pareto chart. Pareto charts are popular in quality control to identify problems in a business process.

▶ If the categorical variable is ordinal, then the bar chart must preserve the ordering.

Example- Pareto chart

A,B,B,C,A,D,B,B,A,C, B,B,C,D,A

Category	Tally mark	Freq	Relative freq
A	||||	4	0.26
B	ⅢⅠ	6	0.40
C	|||	3	0.20
D	||	2	0.14
Total		15	1

Example- ordinal data

The T-shirt sizes (Small-S, Medium-M, Large-L) of twenty students is listed below:

L,M,M,S,L,S,S,M,L,M,M,S,S,L,M,S,M,S,L,M

Size	Tally mark	Freq	Relative freq
Small	||||| ||	7	0.35
Medium	||||| |||	8	0.40
Large	|||||	5	0.25
Total		20	1

Sectional summary

1. A bar chart is used to show the frequencies/relative frequencies of a categorical variable.
2. If ordinal, the order of categories is preserved.
3. The bars can be oriented either horizontally or vertically.
4. A Pareto chart is a bar chart where the categories are sorted by frequency.

Know your purpose

- ▶ Have a purpose for every table or graph you create
  - ▶ Choose the table/graph to serve the purpose.
- ▶ Pie charts are best to use when you are trying to compare parts of a whole.
- ▶ Bar graphs are used to compare things between different groups.

Label your data

- ▶ Label your chart to show the categories and indicate whether some have been combined or omitted.
- ▶ Name the bars in a bar chart.
- ▶ Name the slices in a pie chart.
- ▶ If you have omitted some of the cases, make sure the label of the plot defines the collection that is summarized.

Label your data
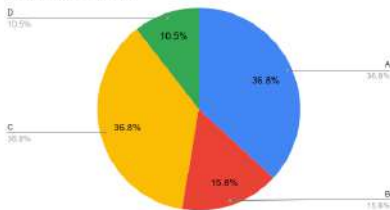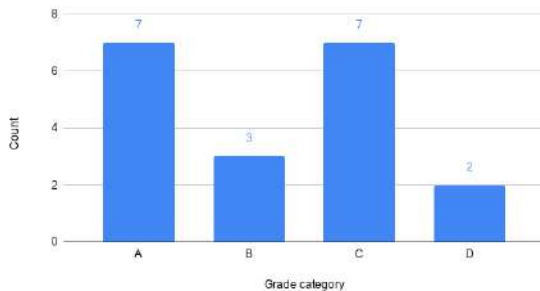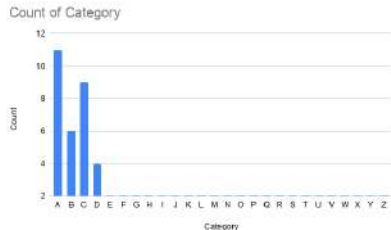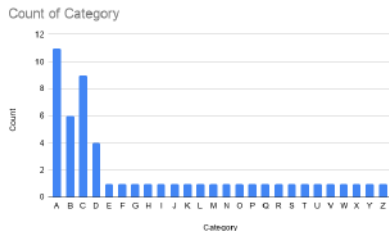
Many categories

A bar chart or pie chart with too many categories might conceal the more important categories. In some case, grouping other categories together might be done.
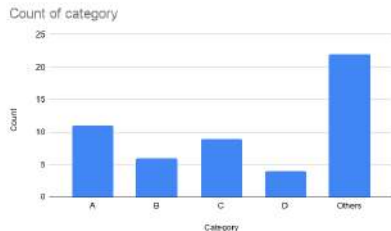
The area principle

- ▶ Displays of data must obey a fundamental rule called the area principle[1].
- ▶ The area principle says that the area occupied by a part of the graph should correspond to the amount of data it represents.
- ▶ Violations of the area principle are a common way to mislead with statistics.

---

[1]Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

Misleading graphs: violating area principle

▶ Decorated graphics: Charts decorated to attract attention often violate the area principle[2]
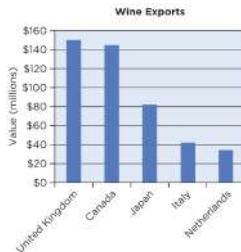
▶ No baseline and the chart shows bottles on top of labeled boxes of various sizes and shapes.

▶ Obeys area principle and accurate

---

[2]Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

Misleading graphs: truncated graphs

▶ Another common violation is when the baseline of a bar chart is not at zero.
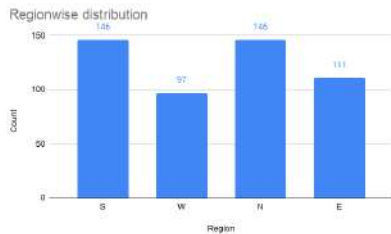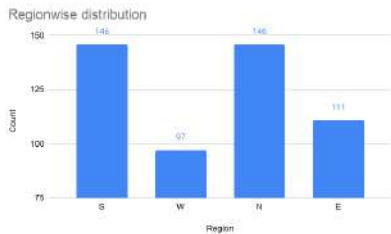
Left graph exaggerates the number coming from the South and North. Graph on right shows same data with the baseline at zero.

Round-off errors

▶ Important to check for round-off errors.

▶ When table entries are percentages or proportions, the total may sum to a value slightly different from 100% or 1. This might result in a pie chart where the total does not add up[3].

---
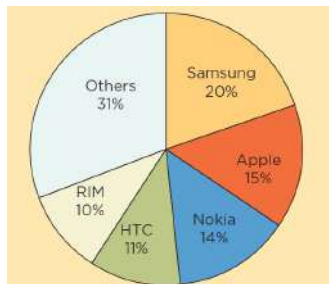
[3]Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

Sectional summary

1. Know your purpose and choose table/graph appropriately
2. Label your charts
3. Handle multiple categories appropriately.
4. Respect area principle
   4.1 Avoid overly decorated graphs
   4.2 Avoid truncated graphs- use special symbols to indicate vertical axis has been modified.
   4.3 Check for round-off errors

Summarizing categorical data

- ▶ Graphical summaries of categorical data: bar chart and pie chart.
- ▶ Need for a compact measure.
- ▶ Numbers that are used to describe data sets are called descriptive measures.
- ▶ Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of central tendency.

Mode

Definition

The mode of a categorical variable is the most common category, the category with the highest frequency

The mode labels

- ▶ The longest bar in a bar chart
- ▶ The widest slice in a pie chart.
- ▶ In a Pareto chart, the mode is the first category shown.

Example

- Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A
- The longest bar in a bar chart
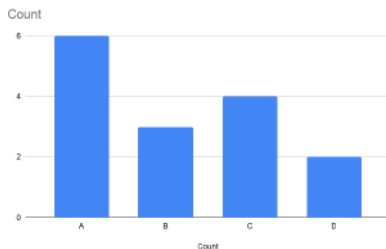
The most common category is "A"

Example

▶ Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A

▶ The widest slice in a pie chart.

Count
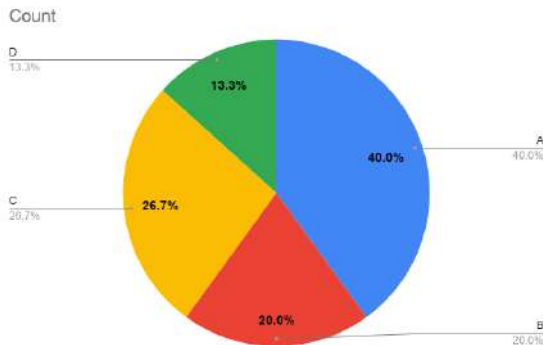
The most common category is "A"

Bimodal and multimodal data

▶ If two or more categories tie for the highest frequency, the data are said to be bimodal (in the case of two) or multimodal (more than two).

▶ Let consider the example A,A,B,C,A,C,A,B,C,C, A,C,C,D,A,A,C,D,B

Count of category
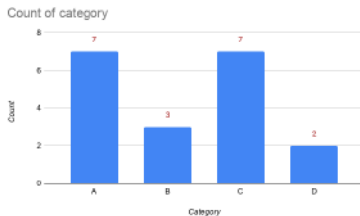
▶ Both category "A" and "C" have highest frequency.

Median

▶ Ordinal data offer another summary, the median, that is not available unless the data can be put into order.

Definition

The median of an ordinal variable is the category of the middle observation of the sorted values.

▶ If there are an even number of observations, choose the category on either side of the middle of the sorted list as the median.

Example

- Consider the grades of 15 students which is listed as
  A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
  - The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D
  - The median grade is the category associated with the 8 observation which is "B".
- Consider the grades of 14 students which is listed as
  A,B,B,C,A,D,B,B,A,C, B,B,C,D
  - The ordered data is A,A,A,B,B,B,B,B,B,C,C,C,D,D
  - The median grade is the category associated with the 7 or 8 observation which is "B".

Example

- ▶ Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A

- ▶ The ordered data is A,A,A,A,B,B,B,B,B,B,C,C,C,D,D

- ▶ The median grade is the category associated with the 8 observation which is "B".

- ▶ The most common grade is "B", hence mode is "B"

- ▶ In this example both mode and median are the same.

Example

▶ Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A

▶ The ordered data is A,A,A,A,A,A,B,B,B,B,C,C,C,D,D

▶ The median grade is the category associated with the 8 observation which is "B".

▶ The most common grade is "A", hence mode is "A"

▶ In this example both mode and median are the different.

Sectional summary

- ▶ The mode of a categorical variable is the most common category.
- ▶ The median of an ordinal variable is the category of the middle observation of the sorted values.

Summary

1. Tabulate data: frequency and relative frequency.
2. Charts of categorical data
   2.1 Pie charts
   2.2 Bar charts and Pareto charts
3. Best practices and misleading graphs
   3.1 Label your data.
   3.2 Dealing with multiple categories.
   3.3 Area principle
   3.4 Misleading graphs
       3.4.1 Decorated graphs
       3.4.2 Truncated graphs.
       3.4.3 Round-off errors.
4. Descriptive measures
   4.1 Mode.
   4.2 Median for ordinal data.

1. **When to use pie chart and bar chart?**
   Use a bar chart to show the frequencies of a categorical variable. Order the categories either alphabetically or by size. The bars can be oriented either horizontally or vertically. Use a pie chart to show the proportions of a categorical variable. Arrange the slices (if you can) to make differences in the sizes more recognizable. A pie chart is a good way to show that one category makes up more than half of the total.
   Consider this situation: A manager has partitioned the company's sales into five cities: Gurgaon, Pune, Mumbai, New Delhi, and Chennai. What graph would you use to make these points in a presentation for management?
   a. A figure that shows that slightly more than half of all sales are made in the Chennai city.
   b. A figure that shows that sales topped 12 crores in every city.
   Use a pie chart for part a. Pie charts emphasize the breakdown of the total into pieces and make it easy to see that more than half of the total sales are in the Chennai city. For part b, use a bar chart because bar charts show the values rather than the relative sizes. Every bar would be long enough to reach past a grid line at 12 crores.

2. **What do you mean by manipulation of y-axis?**
   Expanding or compressing the scale on a graph, that can make changes in the data seem less significant than they actually are, is known as the manipulation of y-axis. For example if we represent the number of sales of smart phone A and B of a local shop, from the first figure we are getting the information that a significant amount of sales is
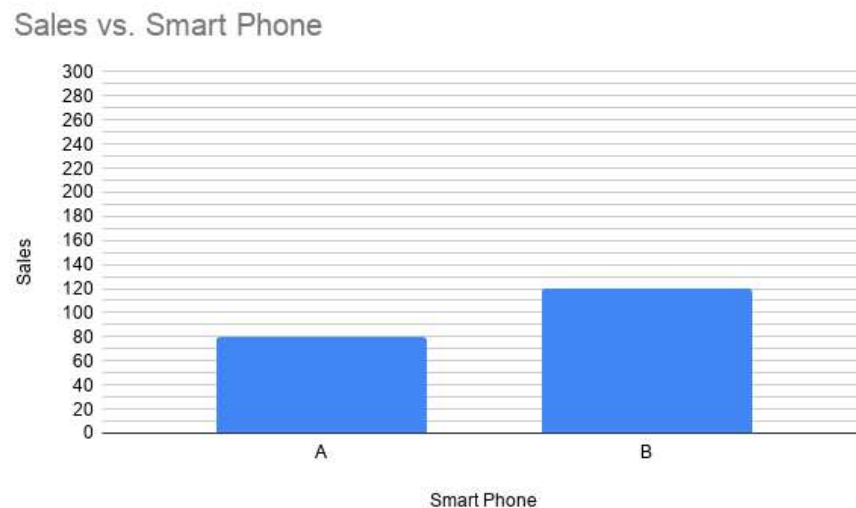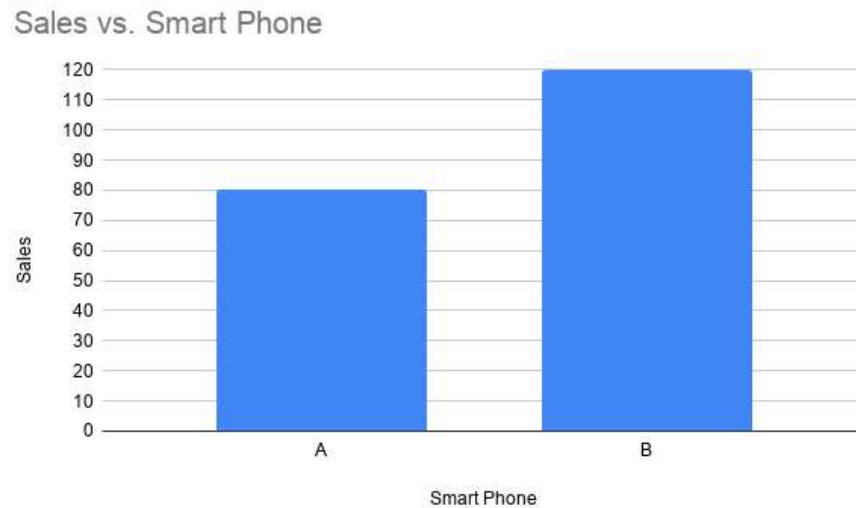
being done of both the smart phones but from the second figure it seems that the sales is very low of the smart phone A and B. So, the second graph is misleading because it has manipulated y-axis.

3. **What do you mean by truncated graph?**
   Omitting baselines, or the axis of a graph, is one of the most common ways data is manipulated in graphs, known as a truncated graph. This misleading tactic is frequently used to make one group look better than another.
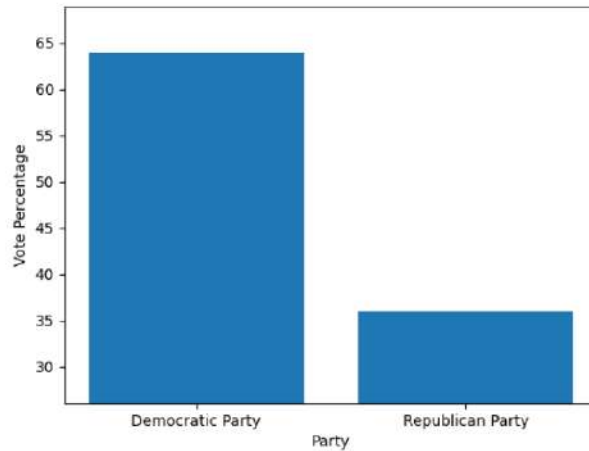

Figure 2.4.5: Share of the votes in an election in the United States of America

For example if we refer the image from AQ 2.4.5, from the length of the bar we observe that Republic party voting percentage is less than half of the Democratic party but if consider the actual number this is not the case.

Review

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
   - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
   - ▶ Understand cross-sectional versus time-series data.
   - ▶ Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - ▶ Creating frequency tables, understanding relative frequency
   - ▶ Creating pie charts and bar charts
   - ▶ Understanding violations
   - ▶ Descriptive measures of Mode and Median

Frequency tables
    Organizing numerical data

Graphical summaries
    Histograms
    Stem-and-leaf diagram

Types of variables
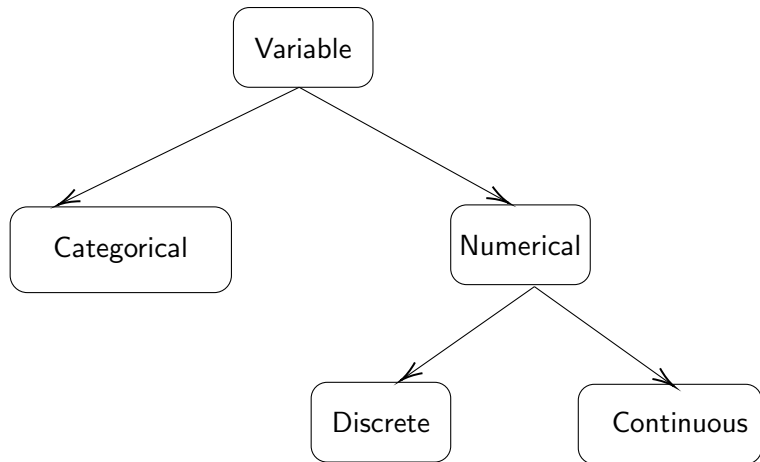
Organizing numerical data

- ▶ Recall, a discrete variable usually involves a count of something, whereas a continuous variable usually involves a measurement of something.

- ▶ First group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of qualitative data.

- ▶ Once we group the quantitative data into classes, we can construct frequency and relative-frequency distributions of the data in exactly the same way as we did for categorical data.

Organizing discrete data (single value)

▶ If the data set contains only a relatively small number of distinct, or different, values, it is convenient to represent it in a frequency table.

▶ Each class represents a distinct value (single value) along with its frequency of occurrence.

Example

▶ Suppose the dataset reports the number of people in a household. The following data is the response from 15 individuals.

▶ 2,1,3,4,5,2,3,3,3,4,4,1,2,3,4

▶ The distinct values the variable, number of people in each household, takes is 1,2,3,4,5.

▶ The frequency distribution table is

Value	Tally mark	Frequency	Relative frequency
1			
2			
3			
4			
5			
Total			

Graph
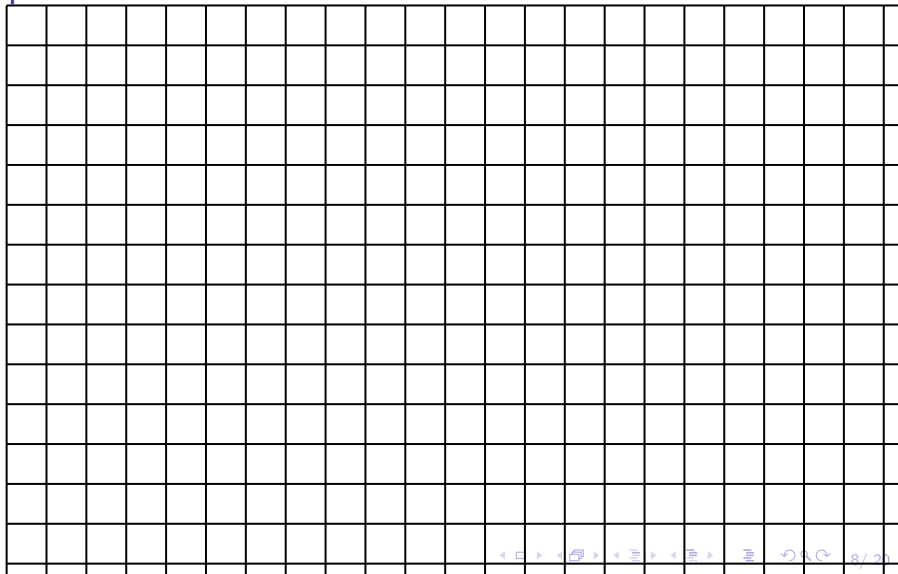
Organizing continuous data

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.

2. Each observation should belong to some class and no observation should belong to more than one class.

3. It is common, although not essential, to choose class intervals of equal length.

Some new terms

1. Lower class limit: The smallest value that could go in a class.
2. Upper class limit: The largest value that could go in a class.
3. Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.
4. Class mark: The average of the two class limits of a class.
5. A class interval contains its left-end but not its right-end boundary point.

Example

- ▶ The marks obtained by 50 students in a particular course.
- ▶ 68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.

Class interval	Tally mark	Frequency	Relative frequency
30-40			
40-50			
50-60			
60-70			
70-80			
80-90			
Total			

Frequency table

68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39,
80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64,
67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.

Class interval	Tally mark	Frequency	Relative frequency
30-40	|||	3	0.06
40-50	ⅢⅠ	6	0.12
50-60	Ⅲ Ⅲ Ⅲ |||	18	0.36
60-70	Ⅲ Ⅲ Ⅲ ||	17	0.34
70-80	||||	4	0.08
80-90	||	2	0.04
Total		50	1

Section summary

1. Frequency table for discrete single value data.
2. Frequency table for continuous data using class intervals.

Steps to construct a histogram

Step 1 Obtain a frequency (relative-frequency) distribution of the data.

Step 2 Draw a horizontal axis on which to place the classes and a vertical axis on which to display the frequencies (relative frequencies).

Step 3 For each class, construct a vertical bar whose height equals the frequency (relative frequency) of that class.

Step 4 Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" ("Relative frequency").

Histogram
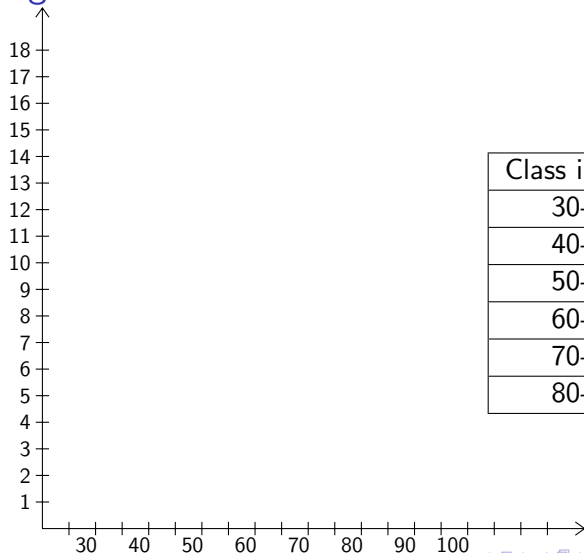
Class interval	frequency
30-40	3
40-50	6
50-60	18
60-70	17
70-80	4
80-90	2

Histogram

```
https://docs.google.com/spreadsheets/d/
109W3ga8TZG3pWJwofG4h0yE7xvoGOK_kCvmmOe9wOkQ/edit?
usp=sharing
```


Distribution of marks
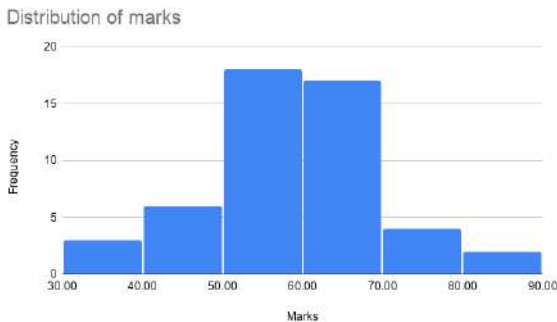
Stem-and-leaf diagram

Definition

In a stem-and-leaf diagram (or stemplot)[1] , each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.

- ▶ For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.
    - ▶ The value 75 is expressed as

Stem		Leaf
7	|	5
    - ▶ The two values 75, 78 is expressed as

Stem		Leaf
7	|	5,8

---

[1]Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.

Steps to construct a stemplot

Step 1 Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

Step 2 Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

Step 3 Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

Step 4 Arrange the leaves in each row in ascending order.

Example

- ▶ The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48
- ▶ Draw a stem-and-leaf plot for this data set.

```
1 | 05
2 | 23589
3 | 16
4 | 58
```

Section summary

1. Construct a histogram for grouped data.
2. Construct a stemplot to describe numerical data.

Descriptive measures

▶ The objective is to develop measures that can be used to summarize a data set.

▶ These descriptive measures are quantities whose values are determined by the data.

Descriptive measures

Most commonly used descriptive measures can be categorized as

▶ Measures of central tendency: These are measures that indicate the most typical value or center of a data set.

▶ Measures of dispersion: These measures indicate the variability or spread of a dataset.

The Mean

The most commonly used measure of central tendency is the mean.

Definition

The mean of a data set is the sum of the observations divided by the number of observations.

▶ The mean is usually referred to as average.

▶ Arithmetic average; divide the sum of the values by the number of values (another typical value)

▶ For discrete observations:

  ▶ Sample mean: $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$

  ▶ Population mean: $\mu = \dfrac{x_1 + x_2 + \ldots + x_N}{N}$

Example

1. 2, 12, 5, 7, 6, 7, 3;
   $\bar{x} = \frac{2+12+5+7+6+7+3}{7} = \frac{42}{7} = 6$
2. 2, 105, 5, 7, 6, 7, 3 $\bar{x} = \frac{2+105+5+7+6+7+3}{7} = \frac{135}{7} = 19.285$
3. 2, 105, 5, 7, 6, 3 $\bar{x} = \frac{2+105+5+7+6+3}{6} = \frac{128}{6} = 21.33$

Example

- The marks obtained by ten students in an exam is
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- The sample mean is

$$\frac{68 + 79 + 38 + 68 + 35 + 70 + 61 + 47 + 58 + 66}{10} = \frac{590}{10} = 59$$

Mean for grouped data: discrete single value data

▶ The following data is the response from 15 individuals.
  2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

▶ $\bar{x} = \dfrac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{n}$

Value(x_i)	Tally mark	Frequency(f_i)	$f_i x_i$
1	‖	2	2
2	‖‖	3	6
3	ЖЖ	5	15
4	‖‖‖	4	16
5	‖	1	5
Total		15	44

Mean$=\frac{44}{15} = 2.93$

Mean for grouped data: continuous data

- $\bar{x} = \dfrac{f_1 m_1 + f_2 m_2 + \ldots + f_n m_n}{n}$

Class interval	Tally mark	Frequency(f_i)	Mid point(m_i)	$f_i m_i$
30-40	|||	3	35	105
40-50	ⅢⅠ	6	45	270
50-60	ⅢⅢ Ⅲ |||	18	55	990
60-70	ⅢⅢ Ⅲ ||	17	65	1105
70-80	||||	4	75	300
80-90	||	2	85	170
Total		50		2940

- Average $= \frac{2940}{50} = 58.8$.

- 58.8 is an <u>approximate</u> and not <u>exact</u> value of the mean

Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- ▶ Example: Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
  - ▶ Suppose the teacher has decided to add 5 marks to each student.
  - ▶ Then the data becomes
    73, 84, 43, 73, 40, 75, 66, 52, 63, 71.
  - ▶ The mean of the new data set is $\frac{640}{10} = 64 = 59 + 5$

Multiplying a constant

▶ Let $y_i = x_i c$ where c is a constant then $\bar{y} = \bar{x} c$

▶ Example: Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66.

  ▶ Suppose the teacher has decided to scale down each mark by
    40%, in other words each mark is multiplied by 0.4.
  ▶ Then the data becomes
    27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
  ▶ The mean of the new data set is $\frac{236}{10} = 23.6 = 59 \times 0.4$

Section summary

1. Mean or average is a measure of central tendency.
2. Compute sample mean for
   2.1 ungrouped data.
   2.2 grouped discrete data.
   2.3 grouped continuous data.
3. Manipulating data
   3.1 Adding a constant to each data point.
   3.2 Multiplying each data point with a constant.

Median

Another frequently used measure of center is the median.
Essentially, the median of a data set is the number that divides the
bottom 50% of the data from the top 50%.

Definition
The median of a data set is the middle value in its ordered list.

Steps to obtain median

Arrange the data in increasing order. Let n be the total number of observations in the dataset.

1. If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e. $\frac{n+1}{2}$ observation

2. If the number of obsevations is even, then the median is the mean of the two middle observations in the ordered list, i.e. mean of $\frac{n}{2}$ and $\frac{n}{2} + 1$ observation

Example

1. 2, 12, 5, 7, 6, 7, 3
   1.1 Arrange the data in increasing order
      2, 3, 5, 6, 7, 7, 12
   1.2 $n = 7$ odd, median is the $\frac{n+1}{2} = \frac{8}{2} = 4^{th}$ observation, "6".
2. 2, 105, 5, 7, 6, 7, 3
   2.1 Arrange the data in increasing order
      2, 3, 5, 6, 7, 7, 105
   2.2 $n = 7$ odd, median is the $\frac{n+1}{2} = \frac{8}{2} = 4^{th}$ observation, "6".
3. 2, 105, 5, 7, 6, 3
   3.1 Arrange the data in increasing order
      2, 3, 5, 6, 7, 105
   3.2 $n = 6$ even, median is the average of $\frac{n}{2}$ and $\frac{n}{2} + 1$ observation
      $= \frac{5+6}{2} = 5.5$.

Example

1. 2, 12, 5, 7, 6, 7, 3
   1.1 Sample mean $= \frac{2+3+5+6+7+7+12}{7} = 6$
   1.2 Sample median $= 6$

2. 2, 117, 5, 7, 6, 7, 3
   2.1 Sample mean $= \frac{2+3+5+6+7+7+117}{7} = 21$
   2.2 Sample median $= 6$

The sample mean is sensitive to outliers, whereas the sample median is not sensitive to outliers.

Adding a constant

▶ Let $y_i = x_i + c$ where c is a constant then
  new median = old median + c

▶ Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  Arranging in ascending order 35,38,47,58,61,66,68,68,70,79
  The median for this data is the average of $\frac{n}{2}$ and $\frac{n}{2} + 1$
  observation which is $\frac{61+66}{2} = \frac{127}{2} = 63.5$

▶ Suppose the teacher has decided to add 5 marks to each
  student.

▶ Then the data in ascending order is
  40,43,52,63,66,71,73,73,75,84

▶ The median of the new dataset is $\frac{66+71}{2} = \frac{137}{2} = 68.5$

▶ Note 68.5=63.5+5

Multiplying a constant

▶ Let $y_i = x_i c$ where c is a constant then

new median = old median $\times$ c

▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66.
We already know median for this data is 63.5

▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

▶ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The ascending order is 14, 15.2, 18.8, 23.2, 24.4, 26.4, 27.2, 28, 31.6
The median of new dataset is $\frac{24.4+26.4}{2} = \frac{50.8}{2} = 25.4$

▶ Note $25.4 = 0.4 \times 63.5$

Mode

Another measure of central tendency is the sample mode.

Definition
The mode of a data set is its most frequently occurring value.

Steps to obtain mode

1. If no value occurs more than once, then the data set has no mode.
2. Else, the value that occurs with the greatest frequency is a mode of the data set.

Example

1. 2, 12, 5, 7, 6, 7, 3;
   7 occurs twice, hence 7 is mode
2. 2, 105, 5, 7, 6, 7, 3
   7 is mode
3. 2, 105, 5, 7, 6, 3 no mode

Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then
  new mode = old mode + c

- ▶ Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66.
  The mode for this data is 68

- ▶ Suppose the teacher has decided to add 5 marks to each
  student.

- ▶ Then the data in ascending order is
  40,43,52,63,66,71,73,73,75,84

- ▶ The mode of the new dataset is 73

- ▶ Note $73 = 68 + 5$

Multiplying a constant

▶ Let $y_i = x_i c$ where c is a constant then

$$new\ mode = old\ mode \times c$$

▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66.
We already know mode for this data is 68

▶ Suppose the teacher has decided to scale down each mark by
40%, in other words each mark is multiplied by 0.4.

▶ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The mode of new dataset is 27.2

▶ Note $27.2 = 0.4 \times 68$

Section summary

- Measures of central tendency
  1. Mean
  2. Median
  3. Mode
- Impact of adding a constant or multiplying with a constant on the measures.

Introduction- why do we need a measure of dispersion

► Consider the two data sets given below
   ► Dataset 1: 3, 3, 3, 3, 3
   ► Dataset 2: 1, 2, 3, 4, 5
► The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2
Mean	3	3
Median	3	3
Mode	3	Not available

► The mean, median are same for both the datasets. However, the datasets are not same. They are different.

Measures of dispersion

▶ To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.

▶ Such descriptive measures are referred to as
  ▶ measures of dispersion, or
  ▶ measures of variation, or
  ▶ measures of spread.

▶ In this course we will be discussing about the following measures of dispersion.
  1. Range.
  2. Variance.
  3. Standard deviation.
  4. Interquartile range.

Range

Definition

The range of a data set is the difference between its largest and smallest values.

▶ The range of a data set is given by the formula

$$\text{Range} = \text{Max - Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

▶
	Dataset 1	Dataset 2
	3,3,3,3,3	1,2,3,4,5
Max	3	5
Min	3	1
Range	0	4

Range sensitive to outliers

▶ Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15
Max	5	15
Min	1	1
Range	4	14

▶ Though the two datasets differ only in one datapoint, we can see that this contributes to the value of Range significantly. This happens because the range takes into consideration only the Min and Max of the dataset.

Variance

▶ In contrast to the Range, the variance takes into account all the observations.

▶ One way of measuring the variability of a data set is to consider the deviations of the data values from a central value

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- ▶ The variance is computed using the following formulae
    - ▶ Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+...+(x_N-\mu)^2}{N}$
    - ▶ Sample variance: $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1}$
- ▶ The numerator is the sum of squared deviations of every observation from its mean.
- ▶ The denominator for computing population variance is N, the total number of observations.
- ▶ The denominator for computing sample variance is $(n-1)$. The reason for this will be clear in forthcoming courses on statistics.

Example

- ▶ Recall marks of students obtained by ten students in an exam is
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- ▶ The mean was computed to be 59.
- ▶ The deviations of each data point from its mean is given in the table below:

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	**590**	**0**	**1898**

1. Population variance $= \frac{1898}{10} = 189.8$

2. Sample variance $= \frac{1898}{9} = 210.88$

Adding a constant
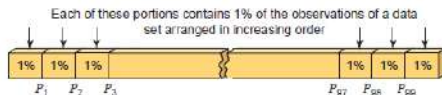
- ▶ Let $y_i = x_i + c$ where c is a constant then
  new variance = old variance
- ▶ Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- ▶ Suppose the teacher has decided to add 5 marks to each student.
- ▶ Then the data is
  73, 84, 43, 73, 40, 75, 66, 52, 63, 71
- ▶ The variance of the new dataset is $\frac{1898}{9} = 210.88$
- ▶ In general, adding a constant does not change variability of a dataset, and hence it is the same.

Multiplying a constant

▶ Let $y_i = x_i c$ where c is a constant then

$$new\ variance = c^2 \times old\ variance$$

▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66.
We already know variance for this data is 210.88

▶ Suppose the teacher has decided to scale down each mark by
40%, in other words each mark is multiplied by 0.4.

▶ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The mean of new dataset is 23.6

▶ The sum of squared deviations from mean $= 303.68$ and the
variance $= \frac{303.68}{9} = 33.74$. We can verify that
$33.74 = 0.4^2 \times 210.88$.

Standard deviation

▶ Another very useful measure of dispersion is the standard
  deviation.

Definition
The quantity

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}}$$

*which is the square root of sample variance is the sample standard
deviation.*

Units of standard deviation

▶ The sample variance is expressed in units of square units if original variable. For example, instead of marks if the data were weights of 10 students measured in kilograms. Then the unit of variance would be $(kilogram)^2$

▶ The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in kilograms, then the units of standard deviation are also in kilograms.

Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then
  new variance = old variance
- ▶ Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- ▶ Suppose the teacher has decided to add 5 marks to each student.
- ▶ Then the data is
  73, 84, 43, 73, 40, 75, 66, 52, 63, 71
- ▶ The variance of the new dataset is $\frac{1898}{9} = 210.88$
- ▶ the standard deviation of the new dataset is
  $\sqrt{210.88} = 14.522$
- ▶ In general, adding a constant does not change variability of a dataset, and hence it is the same.

Multiplying a constant

▶ Let $y_i = x_i c$ where c is a constant then
new variance = $c^2 \times$ old variance

▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66.
We already know variance for this data is 210.88

▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

▶ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The mean of new dataset is 23.6

▶ The sum of squared deviations from mean = 303.68 and the variance = $\frac{303.68}{9} = 33.74$.

▶ The standard deviation of the newdata set is $\sqrt{33.74} = 5.808$. We can verify $5.808 = 0.4 \times 14.522$

Section summary

▶ Measures of dispersion
  1. Range
  2. Variance: population variance and sample variance.
  3. Standard deviation.

▶ Impact of adding a constant or multiplying with a constant on the measures.

Percentiles

▶ The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1 - p)$ percent of the data values are greater than or equal to it.

Each of these portions contains 1% of the observations of a data set arranged in increasing order

| 1% | 1% | 1% | | 1% | 1% | 1% |

P_1 P_2 P_3 P_{97} P_{98} P_{99}

[1]

▶ If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these values.

▶ Median is the 50th percentile.

─────────────────

[1]Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons.

Computing Percentile

To find the sample 100p percentile of a data set of size n

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample 100p percentile.
3. If np is an integer, then the average of the values in positions np and $np + 1$ is the sample 100p percentile.

Example

Let $n = 10$

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68, 68, 70, 79

p	np	
0.1	1	$(35+38)/2 = 36.5$
0.25	2.5	47
0.5	5	$(61+66)/2 = 63.5$
0.75	7.5	68
1	10	79

Computing percentile using googlesheets-PERCENTILE function

Step 1 Paste the dataset in a column.

Step 2 In a blank cell enter PERCENTILE(data, percentile), where data indicates the range of data for which percentile needs to be computed, and percentile is the decimal form of the desired percentile.

▶ For example if the data is in cell A1:A10, and we are interested in computing the 90th percentile, then enter PERCENTILE(A1:A10,0.9) in a blank cell.

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the ith ordered value of the dataset.

Step 2 Find rank using the following formula.
$rank = percentile \times (n-1) + 1$ where n is total number of observations in the dataset

▶ Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10 - 1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

▶ Integer part of $3.25 = 3$; fractional part is 0.25.

Step 4 Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.

▶ The ordered data value corresponding to integer part rank of 3, $x_{[3]}$ is 47.

Computing percentile using googlesheets-algorithm-contd

Step 5 The percentile value is given by the formula

$$Percentile = x_{[i]} + fractional\ part \times \left[x_{[i+1]} - x_{[i]} \right]$$

▶ $Percentile = 47 + 0.25 \times [58 - 47] = 47 + 0.25 \times 11 = 47 + 2.75 = 49.75$

Quartiles

Definition

The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.

In other words, the quartiles break up a data set into four parts with about 25 percent of the data values being less than the first(lower) quartile, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third(upper) quartiles, and about 25 percent being larger than the third quartile.

The Five Number Summary

- ▶ Minimum
- ▶ Q_1: First Quartile or lower quartile
- ▶ Q_2: Second Quartile of Median
- ▶ Q_3: Third Quartile or upper quartile
- ▶ Maximum

The Interquartile Range (IQR)

Definition

The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

.

- ▶ IQR for the example
  - ▶ First quartile, $Q_1 = 49.75$
  - ▶ Third quartile, $Q_3 = 68$
  - ▶ $IQR = Q_3 - Q_1 = 18.25$

Section summary

► Definition of percentiles.

► How to compute percentiles.

► Definition of quartile.

► Five-number summary.

► Interquartile range as a measure of dispersion.

Summary

1. Frequency tables
   1.1 Frequency table for discrete data.
   1.2 Frequency table for continuous data.

2. Graphical summaries
   2.1 Histograms.
   2.2 Stem-an-leaf plot.

3. Numerical summaries
   3.1 Measures of central tendency
       3.1.1 Mean, Median, Mode
   3.2 Measures of dispersion
       3.2.1 Range, Variance, Standard deviation
   3.3 Percentiles
       3.3.1 Interquartile range as a measure of dispersion.

Review

1. What is statistics?
   - ▶ Descriptive statistics, inferential statistics.
2. Understand how data are collected.
   - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
   - ▶ Classify data as categorical or numerical data.
   - ▶ Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - ▶ Creating frequency tables, understanding relative frequency
   - ▶ Creating pie charts and bar charts
   - ▶ Descriptive measures of Mode and Median
5. Describing numerical data
   - ▶ Creating frequency tables: single valued and grouped data.
   - ▶ Measures of central tendency: Mean, Median, and Mode
   - ▶ Measures of dispersion: Range, Variance, Standard deviation
   - ▶ Percentiles, Quartiles, Interquartile range.

Learning objectives

1. Use of two-way contingency tables to understand association between two categorical variables.

2. Understand association between numerical variables through scatter plots; compute and interpret correlation.

3. Understand relationship between a categorical and numerical variable.

Introduction

- ▶ To understand the association between two categorical variables.
- ▶ Learn how to construct two-way contingency table.
- ▶ Learn concept of relative row/column frequencies and how to use them to determine whether there is an association between the categorical variables.

Example 1: Gender versus use of smartphone

▶ A market research firm is interested in finding out whether ownership of a smartphone is associated with gender of a student. In other words, they want to find out whether more females own a smartphone while compared to males, or whether owning a smartphone is independent of gender.

▶ To answer this question, a group of 100 college going children were surveyed about whether they owned a smart phone or not.

▶ The categorical variables in this example are
  ▶ Gender: Male, Female (2 categories)- Nominal variable
  ▶ Own a smartphone: Yes, No (2 categories)- Nominal variable

Example 1: Gender versus use of smartphone-summarize data

- ▶ We have the following summary statistics
    1. There are 44 female and 56 male students
    2. 76 students owned a smartphone, 24 did not own.
    3. 34 female students owned a smartphone, 42 male students owned a smartphone.

- ▶ The data given in the example can be organized using a two-way table, referred to as a contingency table.

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10	34	**44**
Male	14	42	**56**
Column total	**24**	**76**	**100**

Contingency table using google sheets

Step 1 Choose the columns of the variables for which you seek an association.

Step 2 Go to Data-click on Pivot table option

Step 3 Click on create option in the pivot table- it will open the pivot table editor:

    3.1 Under the Rows tab, click on the first categorical variable.

    3.2 Under the columns tab, click on the second categorical variable.

    3.3 Under the values tab, click on either of the variables and then click on the COUNTA tab under "summarize by" tab.

Example 2: Income versus use of smartphone

▶ A market research firm is interested in finding out whether ownership of a smartphone is associated with income of an individual. In other words, they want to find out whether income is associated with ownership of a smartphone.

▶ To answer this question, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not.

▶ The categorical variables in this example are
  ▶ Income: Low, Medium, High (3 categories) -Ordinal variable
  ▶ Own a smartphone: Yes, No (2 categories) - Nominal variable

Example 2: Contingency table

▶ We have the following summary statistics

1. There are 20 High income, 66 medium income, and 14 low income participants.
2. 62 participants owned a smartphone, 38 did not own.
3. 18 High income participants owned a smartphone, 39 Medium income participants owned a smartphone, and 5 Low income participants owned a smartphone.

▶ The contingency table corresponding to the data is given below.

	Own a smartphone		
Income level	**No**	**Yes**	**Row total**
High	2	18	**20**
Medium	27	39	**66**
Low	9	5	**14**
Column total	**38**	**62**	**100**

Section summary

- ▶ Organize bivariate categorical data into a two-way table-contingency table.
- ▶ If data is ordinal, maintain order of the variable in the table

Row relative frequencies

▶ What proportion of total participants own a smart phone?

▶ What proportion of female participants own a smart phone?

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10	34	**44**
Male	14	42	**56**
Column total	**24**	**76**	**100**

Row relative frequency: Divide each cell frequency in a row by its row total.

Example 1: Row relative frequency

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10/44	34/44	**44**
Male	14/56	42/56	**56**
Column total	**24/100**	**76/100**	**100**

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	22.73%	77.27%	**44**
Male	25.00%	75.00%	**56**
Column total	**24.00%**	**76.00%**	**100**

Example 2: Row relative frequency

	Own a smartphone		
Income level	**No**	**Yes**	**Row total**
High	2/20	18/20	**20**
Medium	27/66	39/66	**66**
Low	9/14	5/14	**14**
Column total	**38/100**	**62/100**	**100**

	Own a smartphone		
Income level	**No**	**Yes**	**Row Total**
High	10.00%	90.00%	**20**
Medium	40.91%	59.09%	**66**
Low	64.29%	35.71%	**14**
Column Total	**38.00%**	**62.00%**	**100**

Column relative frequencies

► What proportion of total participants are female?

► What proportion of smart phone owners are females?

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10	34	**44**
Male	14	42	**56**
Column total	**24**	**76**	**100**

Column relative frequency: Divide each cell frequency in a column by its column total.

Example 1: Column relative frequency

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10/24	34/76	**44/100**
Male	14/24	42/76	**56/100**
Column total	**24**	**76**	**100**

	Own a smartphone		
Gender	No	Yes	**Row Total**
Female	41.67%	44.74%	**44.00%**
Male	58.33%	55.26%	**56.00%**
Column Total	**24**	**76**	**100**

Example 2: Column relative frequency

	Own a smartphone		
Income level	**No**	**Yes**	**Row total**
High	2/38	18/62	**20/100**
Medium	27/38	39/62	**66/100**
Low	9/38	5/62	**14/100**
Column total	**38**	**62**	**100**

	Own a smartphone		
Income level	**No**	**Yes**	**Row Total**
High	5.26%	29.03%	**20.00%**
Medium	71.05%	62.90%	**66.00%**
Low	23.68%	8.06%	**14.00%**
Column Total	**38**	**62**	**100**

Section summary

▶ Concept of relative frequency: row relative frequency and column relative frequency.

Association between two variables

▶ What do we mean by stating two variables are associated?
  Knowing information about one variable provides information about the other variable.

▶ To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies described earlier.

Association between two variables

- ▶ If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.

- ▶ If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.

Example 1: Association between two variables

► If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.

Gender	Own a smartphone		Row total
	No	Yes	
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

Gender	Own a smartphone		Row Total
	No	Yes	
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100

Gender and smartphone ownership are not associated

Example 2: Association between two variables

▶ If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.

Income level	Own a smartphone No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Income level	Own a smartphone No	Yes	Row Total
High	5.26%	29.03%	20.00%
Medium	71.05%	62.90%	66.00%
Low	23.68%	8.06%	14.00%
Column Total	38	62	100

Income and smartphone ownership are associated

Stacked bar chart

▶ Recall, a bar chart summarized the data for a categorical variable. It presented a graphical summary of the categorical variable under consideration, with the length of the bars representing the frequency of occurrence of a particular category.

▶ A stacked bar chart represents the counts for a particular category. In addition, each bar is further broken down into smaller segments, with each segment representing the frequency of that particular category within the segment. A stacked bar chart is also referred to as a segmented bar chart.

Stacked bar chart using google sheets

Step 1: Select the data you want to include in the contingency table.

Step 2: Click Insert - chart- choose stacked bar option

Example 1: Stacked bar chart

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	22.73%	77.27%	**44**
Male	25.00%	75.00%	**56**
Column total	**24.00%**	**76.00%**	**100**

Gender versus smartphone ownership
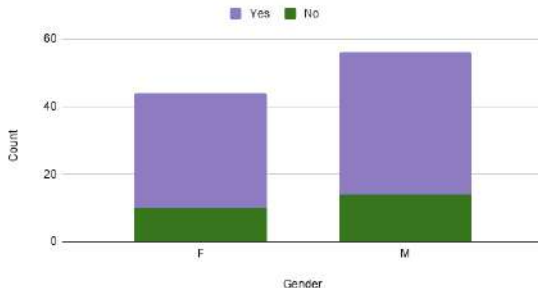
Example 1: 100% Stacked bar chart

A 100% stacked bar chart is useful to part-to-whole relationships
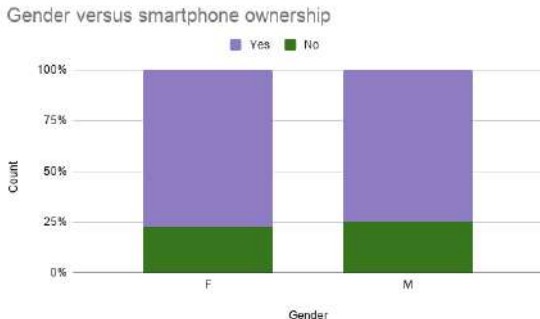
Example 2: Stacked bar chart

	Own a smartphone		
Income level	No	Yes	Row Total
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Income versus smartphone ownership
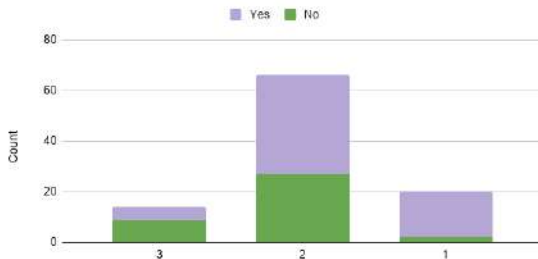
Example 2: 100% Stacked bar chart
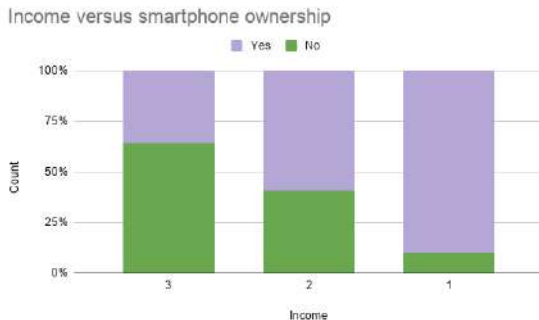
Section summary

▶ Understand whether two categorical variables are associated using the concept of relative frequencies.

▶ Graphical summary of association using stacked bar chart.

Introduction

- ▶ To understand the association between two numerical variables.
- ▶ Learn how to construct scatter plots and interpret association in scatter plots.
- ▶ Summarize association with a line.
- ▶ Correlation matrix.

Scatter plot

We use a scatterplot to look for association between numerical variables.

Definition

A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane.

▶ To decide which variable to put on the x-axis and which to put on the y-axis, display the variable you would like to explain along the y-axis (referred as response variable) and the variable which explains on x-axis (referred as explanatory variable).

Example 1

Age (years)	Height (cms)
1	75
2	85
3	94
4	101
5	108
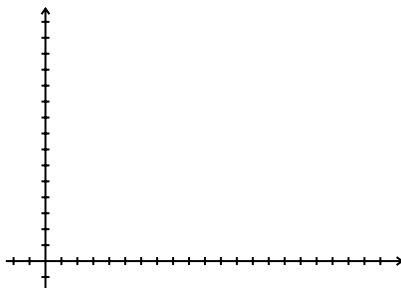
Example 2: Prices of homes

A real estate agent collected the prices of different sizes of homes. He wanted to see what was the relationship between the price of a home and size of a home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way? To answer the question, he collected data on 15 homes. The data he recorded was

1. Size of a home measured in 1000 of square feet.
2. Price of a home measured in lakh of rupees.

Housing data

	Size (1000 Square feet)	Price (INR Lakhs)
1	0.8	68
2	1	81
3	1.1	72
4	1.3	91
5	1.6	87
6	1.8	56
7	2.3	83
8	2.3	112
9	2.5	93
10	2.5	98
11	2.7	136
12	3.1	109
13	3.1	122
14	3.2	159
15	3.4	170

Scatter plot using google sheets

Step 1: Highlight data you want to plot

Step 2: Insert - chart- choose scatter chart

Step 3: Under $X-$axis tab, choose your explanatory variable.

Step 4: Under series tab, the response variable.

Step 5: Label the title of the chart, axes appropriately.

Scatter plot

Visual test for association

▶ Do we see a pattern in the scatter plot?

  ▶ In other words, if I know about the x-value, can I use it to say
    something about the y-value or guess y-value?
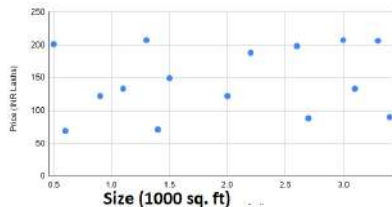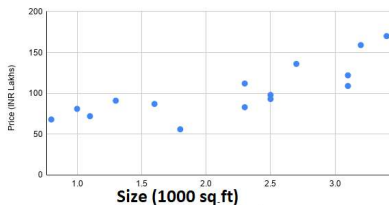
Section summary

1. Draw a scatter plot
2. Notion of explanatory variable and response variable.
3. Visual test for association

Describing association

When describing association between varaibles in a scatter plot, there are four key questions[1] that need to be answered

1. Direction: Does the pattern trend up, down, or both?

2. Curvature: Does the pattern appear to be linear or does it curve?

3. Variation: Are the points tightly clustered along the pattern?

4. Outliers: Did you find something unexpected?

---

[1]Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

Describing association: Curvature

Does the pattern appear to be linear or does it curve?
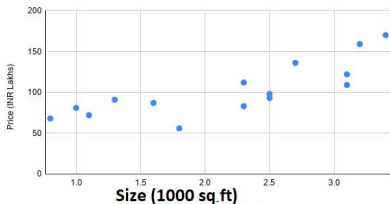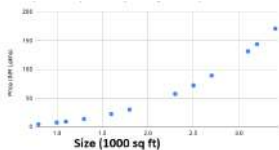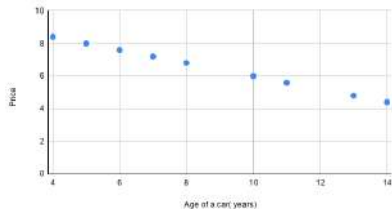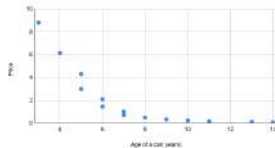
Describing association: Variation

Are the points tightly clustered along the pattern?
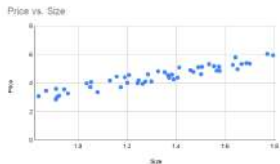
Tightly clustered
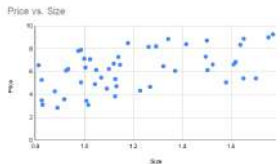
Variable

Describing association: Outliers

Did you find something unexpected?

Price vs. Size
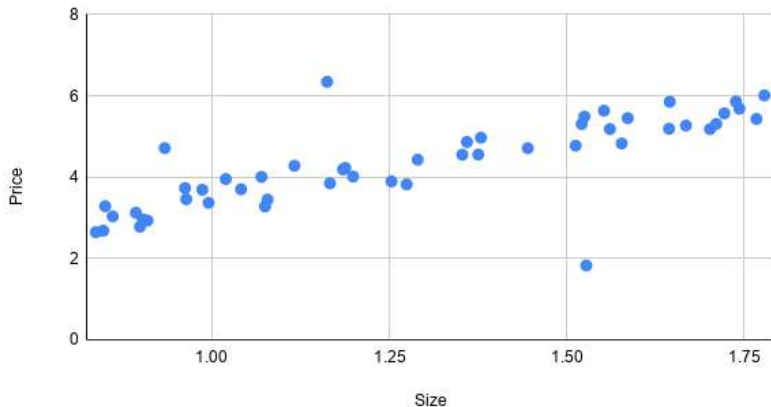
Section summary

Describing association

1. Direction

2. Curvature

3. Variation

4. Outliers.

Measures of association

How do we measure the strength of association between two
variables?

1. Covariance
2. Correlation

Covariance

Covariance quantifies the strength of the linear association
between two numerical variables.

Covariance: Example 2

Variables: Age of a car and price of a car

Age (years)	Price (INR lakhs)	Deviation of x	Deviation of y
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$
1	6	-2	2
2	5	-1	1
3	4	0	0
4	3	1	-1
5	2	2	-1
3	**4**		

Covariance: Example 1

Age	Height	Deviation of x	Deviation of y	
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8

Covariance: Example 2

Age	Price	Deviation of x	Deviation of y	
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4

Key observation

▶ When large (small) values of x tend to be associated with large (small) values of y- the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be same.

▶ When large (small) values of x tend to be associated with small (large) values of y- the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

Covariance

Definition
Let x_i denote the ith observation of variable x, and y_i denote the ith observation of variable y. Let (x_i, y_i) be the ith paired observation of a population (sample) dataset having $N(n)$ observations. The Covariance between the variables x and y is given by

▶ *Population covariance:* $Cov(x, y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N}$

▶ *Sample covariance:* $Cov(x, y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Covariance: Example 1

Age	Height	Deviation of x	Deviation of y	
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8
				82

▶ Population covariance: $\frac{82}{5} = 16.4$

▶ Sample covariance: $\frac{82}{4} = 20.5$

Covariance: Example 2

Age	Price	Deviation of x	Deviation of y	
x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
				-10

▶ Population covariance: $\frac{-10}{5} = -2$

▶ Sample covariance: $\frac{-10}{4} = -2.5$

Units of Covariance

▶ The size of the covariance, however, is difficult to interpret because the covariance has units.

▶ The units of the covariance are those of the x-variable times those of the y-variable.

Section summary

1. Introduced the measure of covariance
2. How to interpret the covariance measure

Learning objectives

1. Understand the measure of correlation.
2. Interpret correlation to quantify the strength of association between two numerical variables.

Correlation

▶ A more easily intepreted measure of linear association between two numerical variables is correlation

▶ It is derived from covariance.

▶ To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y. The Pearson correlation coefficient, r, between x and y is given by

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

Remark
The units of the standard deviations cancel out the units of covariance

Remark
It can be shown that the correlation measure always lies between -1 and +1

Correlation: Example 1

Age	Height	sq.Devn of x	sq.Devn of y	
x	y	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	4	309.76	35.2
2	85	1	57.76	7.6
3	94	0	1.96	0
4	101	1	70.56	8.4
5	108	4	237.16	30.8
		10	**677.2**	**82**

- $s_x = 1.58$, $s_y = 13.01$
- $r = \frac{82}{\sqrt{10 \times 677.2}}$ OR $\frac{20.5}{1.58 \times 13.01} = 0.9964$

Correlation: Example 2

Age	Price	sq. Devn of x	sq. Devn of y	
x	y	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	4	4	-4
2	5	1	1	-1
3	4	0	0	0
4	3	1	1	-1
5	2	4	4	-4
		10	**10**	**-10**

▶ $s_x = 1.58$, $s_y = 1.58$

▶ $r = \frac{-10}{\sqrt{10} \times \sqrt{10}}$ OR $\frac{-2.5}{1.58 \times 1.58} = -1$

Correlation using google sheets

Step 1 The function CORREL(series1, series2) will return the value of correlation.

For example: If the data corresponding to x-variable (series1) is in cell A2:A6 and data corresponding to y-variable (series2) is in cells B2:B6; then CORREL(A2:A6,B2:B6) returns the value of the Pearson Correlation coefficient.

Section summary

1. Introduced measure of correlation.
2. Interpreting correlation between variables.

Learning objectives

1. Summarize the linear association between two variables using the equation of a line.
2. Understand the significance of R^2

Summarizing the association with a line

▶ The strength of linear association between the variables was measured using the measures of Covariance and Correlation.

▶ The linear association can be described using the equation of a line.

Equation of line using google sheets

Step 1 Open the scatter plot

Step 2 Under customize tab, click on series

Step 3 Click on trendline

Step 4 Under label tab, click on use equation, and click the show R^2 button.

Statistics for Data Science -1
└─ Association between numerical variables
  └─ Fitting a line

Example 1: Size versus Price of homes: Equation

Price (INR Lakhs) vs. Size (1000 Square feet)
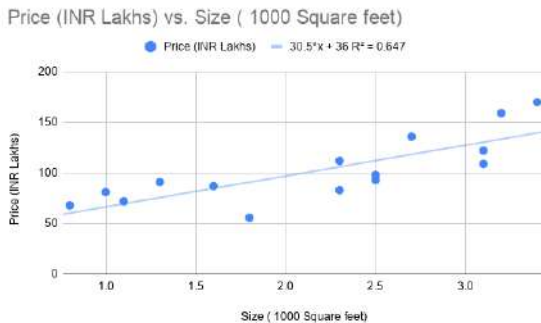
Equation of the line: $Price = 30.5 \times Size + 36$;
$R^2 = 0.647; r = 0.804$

Statistics for Data Science -1
└─ Association between numerical variables
  └─ Fitting a line

Example 2: Age versus Price of cars: Equation

Price vs. Age of a car(years)
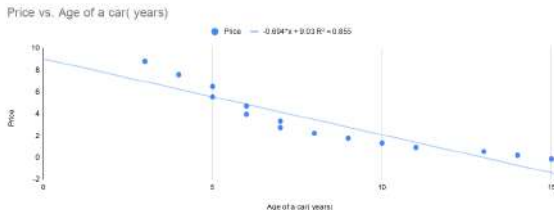
Equation of the line: $Price = -0.694 \times Age + 9.03$;
$R^2 = 0.855; r = -0.9247$

Statistics for Data Science -1
  └─Association between numerical variables
     └─Fitting a line

Example 3: Size versus Price of homes: Equation

Price (INR Lakhs) vs. Size (1000 Square feet)
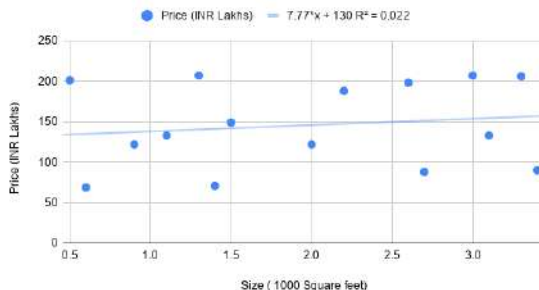
Equation of the line: $Price = 7.77 \times Size + 130$;
$R^2 = 0.022; r = 0.149$

Statistics for Data Science -1
└─ Association between numerical variables
  └─ Fitting a line

Section summary

1. Equation of a line describing linear relationship between two variables.

2. Interpreting slope, R^2 of the line.

Introduction

▶ Understand the association between a categorical variable and numerical variable.

▶ Assume the categorical variable has two categories (dichotomous)

Example 1: Gender versus marks

A teacher was interested in knowing if female students performed better than male students in her class. She collected data from twenty students and the marks they obtained on 100 in the subject.

Example 1: Gender versus marks-Data

	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90

Example 1: Scatter plot

Gender-coded and Marks

Example 1: Scatter plot

Gender-coded and Marks-2

Point Bi-serial Correlation Coefficient

▶ Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).

▶ The following steps are used for calculating the Point Bi-serial correlation between these two variables:

Step 1 Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.

Step 2 Calculate the mean values of two groups: Let $\bar{Y}_0$ and $\bar{Y}_1$ be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

Step 3 Let p_0 and p_1 be the proportion of observations in a group with $Y = 0$ and $Y = 1$, respectively, and s_X be the standard deviation of the random variable X.

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{s_x} \right) \sqrt{p_0 p_1}$$

Learning objectives

1. Understand basic principles of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.

Statistics for Data Science -1
└─ Basic principles of counting
  └─ Addition rule of counting

Example 1: Buying clothes

▶ You have a gift card from a major retailer which allows you to buy "one" item, either a shirt **or** a pant.

▶ The choices at the retailer are

4   3

▶ How many different ways can you use your card?

Statistics for Data Science -1
└─ Basic principles of counting
  └─ Addition rule of counting

Solution

- ▶ There are four choices for buying a shirt
- ▶ There are three choices for buying a pant
- ▶ If you choose to buy a shirt (pant), you cannot buy a pant (shirt).
- ▶ Hence, the total choices available are $4 + 3 = 7$

Addition rule of counting

▶ If an action A can occur in n_1 different ways, another action B can occur in n_2 different ways, then the total number of occurrence of the actions A **or** B is $n_1 + n_2$.

Example 2: Matching shirts and pants

- ▶ Suppose now your card allows you to buy one shirt **and** one pant- how many choices do you have?

- ▶ Suppose we have four shirts and three pants. How many sets can we make?

Matching shirts and pants

4   3

3   3   3   3

Tree
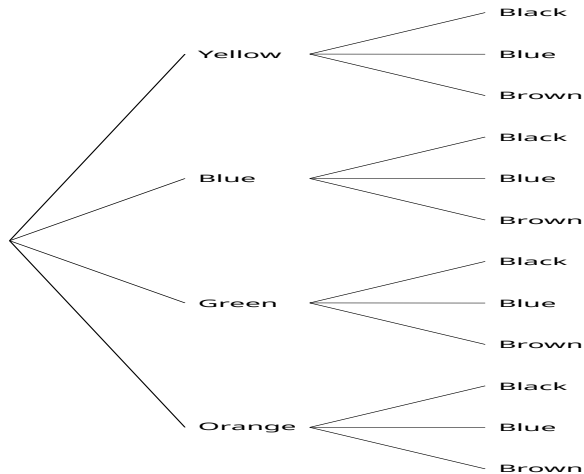
Matching shirts and pants and shoes

Statistics for Data Science -1
└─ Basic principles of counting
    └─ Multiplication rule of counting
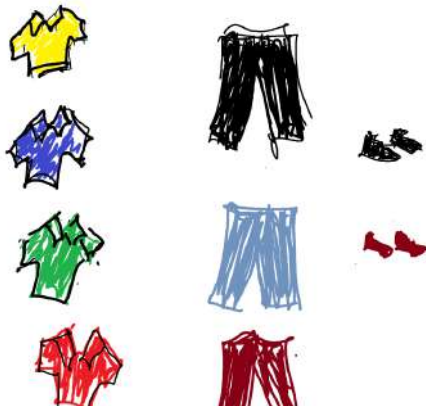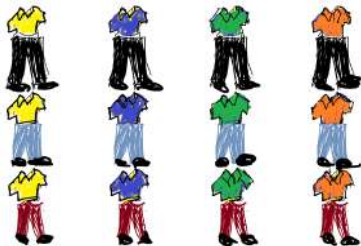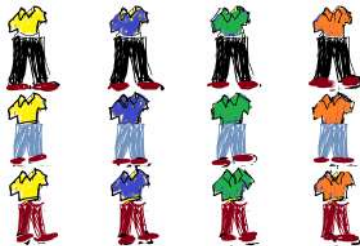
12 ways   12 ways

Total 12+12= 24 ways

Tree
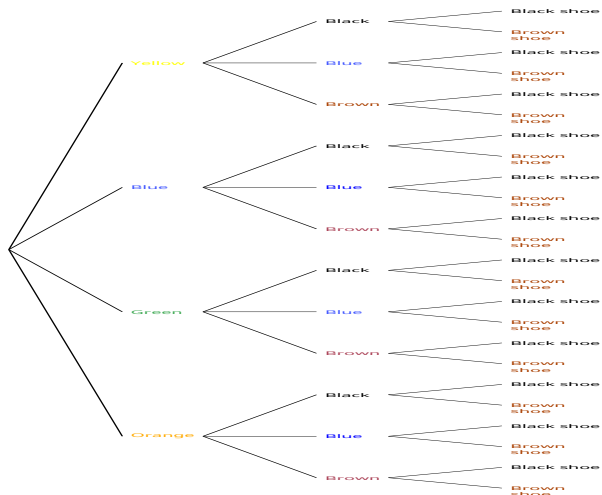
Multiplication rule of counting

▶ If an action A can occur in n_1 different ways, another action B can occur in n_2 different ways, then the total number of occurrence of the actions A **and** B together is $n_1 \times n_2$.

▶ Suppose that r actions are to be performed in a definite order. Further suppose that there are n_1 possibilities for the first action and that corresponding to each of these possibilities are n_2 possibilities for the second action, and so on. Then there are $n_1 \times n_2 \times \ldots \times n_r$ possibilities altogether for the r actions.

Statistics for Data Science -1
└─ Basic principles of counting
  └─ Multiplication rule of counting

Example 2: Application: Creating alpha-numeric code

▶ Suppose you are asked to create a six digit alpha-numeric password with the following requirement:

▶ The password should have first two letters followed by four numbers.

▶ Repetition allowed.
  ▶ Number of ways- $26 \times 26 \times 10 \times 10 \times 10 \times 10 = 6,760,000$

▶ Repetition not allowed.
  ▶ Number of ways- $26 \times 25 \times 10 \times 9 \times 8 \times 7 = 3,276,000$

Statistics for Data Science -1
└─ Basic principles of counting
  └─ Multiplication rule of counting

Section summary

- ▶ Addition rule of counting.
- ▶ Multiplication rule of counting.

Example 3: Order of finishes in a race

- ▶ There are eight athletes who take part in a 100 m race. What are the possible ways the athletes can finish the race (assuming no ties)?

- ▶ First place - any one of the 8 athletes; second - any one of the remaining 7, and so on, the seventh place - any one of the remaining 2, and finally the last place goes to the only one remaining.

- ▶ Hence the total number of ways $=$
  $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 40,320$

Factorial

Definition
The product of the first n positive integers (counting numbers) is called n factorial and is denoted n!. In symbols,

$$n! = n \times (n-1) \times \ldots \times 1$$

Remark
By convention $0! = 1$

Example 4: Choosing shirts
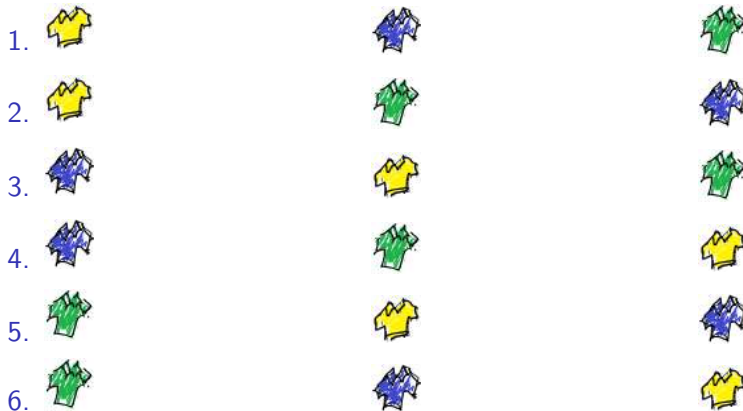
Example 5

1. $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$
2. Observe $5! = 5 \times 4!$
   - In general,
   $$n! = n \times (n-1)!$$
3. Observe $5! = 5 \times 4! = 5 \times 4 \times 3!$
   - In general, for $i \leq n$ we have,
   $$n! = n \times (n-1)\ldots \times (n-i+1) \times (n-i)!$$

Example 6: Simplifying expressions

1. $\frac{6!}{3!} = \frac{6 \times 5 \times 4 \times 3!}{3!} = 6 \times 5 \times 4 = 120$
2. $\frac{6! \times 5!}{3! \times 4!} = \frac{6 \times 5 \times 4 \times 3!}{3!} \frac{5 \times 4!}{4!} = 6 \times 5 \times 4 \times 5 = 600$
3. Express $25 \times 24 \times 23$ in terms of factorials-

$$\frac{25 \times 24 \times 23 \times 22 \times \ldots \times 1}{22 \times 21 \times \ldots \times 1} = \frac{25!}{22!}$$

Section summary

▶ Introduced factorial notation.
▶ Simplifying expressions.

Learning objectives

1. Understand basic principles of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.

Learning objectives

1. Understand basic principles of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.

Permutations

Permutation when objects are distinct

Permutation when objects are distinct- repetitions not allowed

Permutation when objects are distinct- repetitions allowed

Permutation

Definition

A permutation is an ordered arrangement of all or some of n objects.

Example

Take A, B, C- Possible arrangements- taking all at a time

First place	Second place	Third place
A	B	C
A	C	B
B	A	C
B	C	A
C	A	B
C	B	A

Example

Take A, B, C- Possible arrangements- taking two at a time

First place	Second place
A	B
A	C
B	A
B	C
C	A
C	B

Example

Take A, B, C, D- Possible arrangements- taking all at a time

First place	Second place	Third place	Fourth place
A	B	C	D
A	B	D	C
A	C	B	D
A	C	D	B
A	D	B	C
A	D	C	B
B	A	C	D
B	A	D	C
B	C	A	D
B	C	D	A
B	D	A	C
B	D	C	A
C	A	B	D
C	A	D	B
C	B	A	D
C	B	D	A
C	D	A	B
C	D	B	A
D	A	B	C
D	A	C	B
D	B	A	C
D	B	C	A
D	C	A	B
D	C	B	A

Example

Take A, B, C, D- Possible arrangements- taking two at a time

First place	Second place
A	B
A	C
A	D
B	A
B	C
B	D
C	A
C	B
C	D
D	A
D	B
D	C

Permutation formula

The number of possible permutations of r objects from a collection of n distinct objects is given by the formula

$$n \times (n-1) \times \ldots \times (n-r+1)$$

and is denoted by ${}^{n}P_{r}$

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

▶ Special cases

  1. ${}^{n}P_{0} = \frac{n!}{(n-0)!} = \frac{n!}{n!} = 1$ There is only one ordered arrangement of 0 objects.
  2. ${}^{n}P_{1} = \frac{n!}{(n-1)!} = n$. There are n ways of choosing one object from n objects.
  3. ${}^{n}P_{n} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$. We can arrange n distinct objects in $n!$ ways- multiplication principle of counting.

Permutation formula

The number of possible permutations of r objects from a collection of n distinct objects is given by the formula

$$n \times (n-1) \times \ldots \times (n-r+1)$$

and is denoted by ^nP_r

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

▶ Special cases

1. $^nP_0 = \frac{n!}{(n-0)!} = \frac{n!}{n!} = 1$ There is only one ordered arrangement of 0 objects.
2. $^nP_1 = \frac{n!}{(n-1)!} = n$. There are n ways of choosing one object from n objects.
3. $^nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$. We can arrange n distinct objects in $n!$ ways- multiplication principle of counting.

Example

Take A, B, C- Possible arrangements- taking all at a time

First place	Second place	Third place
A	B	C
A	C	B
B	A	C
B	C	A
C	A	B
C	B	A

$$n = 3, r = 3, \, {}^{n}P_{r} = \frac{n!}{(n-r)!} = \frac{3!}{0!} = 6$$

Example

Take A, B, C- Possible arrangements- taking two at a time

First place	Second place
A	B
A	C
B	A
B	C
C	A
C	B

$$n = 3, r = 2, {}^nP_r = \frac{n!}{(n-r)!} = \frac{3!}{1!} = 6$$

Example

Take A, B, C, D- Possible arrangements- taking all at a time

First place	Second place	Third place	Fourth place
A	B	C	D
A	B	D	C
A	C	B	D
A	C	D	B
A	D	B	C
A	D	C	B
B	A	C	D
B	A	D	C
B	C	A	D
B	C	D	A
B	D	A	C
B	D	C	A
C	A	B	D
C	A	D	B
C	B	A	D
C	B	D	A
C	D	A	B
C	D	B	A
D	A	B	C
D	A	C	B
D	B	A	C
D	B	C	A
D	C	A	B
D	C	B	A

$$n = 4, r = 4, \ ^nP_r = \frac{n!}{(n-r)!} = \frac{4!}{0!} = 24$$

Example

Take A, B, C, D- Possible arrangements- taking two at a time

First place	Second place
A	B
A	C
A	D
B	A
B	C
B	D
C	A
C	B
C	D
D	A
D	B
D	C

$$n = 4, r = 2, {}^nP_r = \frac{n!}{(n-r)!} = \frac{4!}{2!} = 12$$

Example: application

▶ From a committee of 8 persons, in how many ways can we choose a chairman and a vice chairman assuming one person can not hold more than one position?

▶ $8 \times 7 = 56$

Example: application

- ▶ Find the number of 4-digit numbers that can be formed using the digits 1, 2, 3, 4, 5 if no digit is repeated.
- ▶ $5 \times 4 \times 3 \times 2 \times 1 = 120$
- ▶ How many of these will be even? 48

Example: application

- ▶ Six people go to the cinema. They sit in a row with ten seats. Find how many ways can this be done if
    - ▶ (i) they can sit anywhere: $^{10}P_6 = 1,51,200$
    - ▶ (ii) all the empty seats are next to each other: $^{7}P_6 = 5,040$

Example

Take A, B, C- Possible arrangements- taking all at a time

First place	Second place	Third place
A	A	A
A	A	B
A	A	C
A	B	A
A	B	B
A	B	C
A	C	A
A	C	B
A	C	C
B	A	A
B	A	B
B	A	C
B	B	A
B	B	B
B	B	C
B	C	A
B	C	B
B	C	C
C	A	A
C	A	B
C	A	C
C	B	A
C	B	B
C	B	C
C	C	A
C	C	B
C	C	C

Example

Take A, B, C- Possible arrangements- taking two at a time

First place	Second place
A	A
A	B
A	C
B	A
B	B
B	C
C	A
C	B
C	C

Permutation formula

The number of possible permutations of r objects from a collection of n distinct objects when repetition is allowed is given by the formula

$$n \times n \times \ldots \times n$$

and is denoted by n^r

Example

Take A, B, C- Possible arrangements- taking all at a time. $n = 3$, $r = 3$, $n^r = 27$

First place	Second place	Third place
A	A	A
A	A	B
A	A	C
A	B	A
A	B	B
A	B	C
A	C	A
A	C	B
A	C	C
B	A	A
B	A	B
B	A	C
B	B	A
B	B	B
B	B	C
B	C	A
B	C	B
B	C	C
C	A	A
C	A	B
C	A	C
C	B	A
C	B	B
C	B	C
C	C	A
C	C	B
C	C	C

Example

Take A, B, C- Possible arrangements- taking two at a time

First place	Second place
A	A
A	B
A	C
B	A
B	B
B	C
C	A
C	B
C	C

$n = 3, r = 2, n^r = 9$

Section summary

1. The number of possible permutations of r objects from a collection of n distinct objects is given by the formula

$$n \times (n-1) \times \ldots \times (n-r+1)$$

and is denoted by ${}^nP_r = \frac{n!}{(n-r)!}$

2. The number of possible permutations of r objects from a collection of n distinct objects when repetition is allowed is given by the formula

$$n \times n \times \ldots \times n$$

and is denoted by n^r

Learning objectives

1. Understand basic principles of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.

Permutations

Permutation when objects are not distinct

Circular permutations

Solving of n and r using permutation formula

Example: Rearranging letters

▶ Suppose we want to rearrange the letters in the word "DATA". How many ways can it be done?

▶ There are three distinct letters : D, A, T.

▶ Hence the possible arrangements taking all the four letters at a time are

First place	Second place	Third place	Fourth place
A	D	T	A
A	D	A	T
A	T	D	A
A	T	A	D
A	A	D	T
A	A	T	D
D	A	T	A
D	A	A	T
D	T	A	A
T	A	D	A
T	A	A	D
T	D	A	A

Permutations when objects are not distinct

▶ As seen in the example, we can treat the two A's in DATA as distinct. Say, A_1 and A_2.

▶ If they are treated as distinct objects, then based on the earlier formula, total number of arrangements $= 4!$.

▶ Now A_1 and A_2 can be arranged among themselves in $2!$ ways.

▶ A_1 and A_2 are essentially the same. Hence, the total number of ways the letters in "DATA" can be arranged is $\frac{4!}{2!} = 12$

Permutation formula

▶ The number of permutations of n objects when p of them are of one kind and rest distinct is equal to

$$\frac{n!}{p!}$$

Example

- ▶ Suppose we want to rearrange the letters in the word "STATISTICS". How many ways can it be done?
- ▶ Total of ten letters of which there are five distinct letters : S,T,A,I,C.
- ▶ "S" appears 3 times; "T" appears 3 times, "A' once, "I" twice, and "C" once

.

Permutation formula

▶ The number of permutations of n objects where p_1 is of one kind, p_2 is of second kind, and so on p_k of kth kind is given by

$$\frac{n!}{p_1! p_2! \ldots p_k!}$$

▶ Applying the above formula to the word " STATISTICS"; $n = 10, p_1 = 3, p_2 = 3, p_3 = 1, p_4 = 2, p_5 = 1$. Hence, total number of ways $=$

$$\frac{10!}{3!3!1!2!1!} = 50,400$$

Section summary

1. The number of permutations of n objects when p of them are of one kind and rest distinct is equal to $\frac{n!}{p!}$

2. The number of permutations of n objects where p_1 is of one kind, p_2 is of second kind, and so on p_k of kth kind is given by $\frac{n!}{p_1! p_2! ... p_k!}$

Example

- ▶ How many ways can four people sit in a round table?
- ▶ We consider two cases: each selection is called a combination of 3 different objects taken 2 at a time.
    - ▶ Clockwise and anticlockwise are different
    - ▶ Clockwise and anticlockwise are same.

Circular permutation: Clockwise and anticlockwise are different

- ▶ Consider the linear permutations of A, B, C and D
- ▶ The arrangements $ABCD$, $BCDA$, $CDAB$, and $DABC$ are different when the people are seated in a row.
- ▶ However, when they are seated in a circle as shown below:
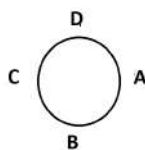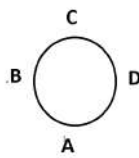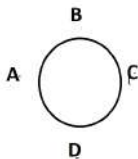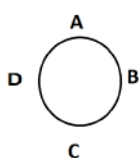
Circular permutation: Clockwise and anticlockwise are different
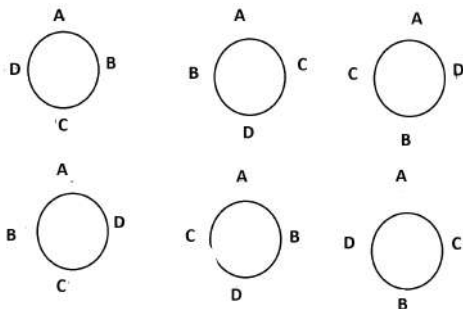
The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are different) is equal to $(n-1)!$

Circular permutation: Clockwise and anticlockwise are same
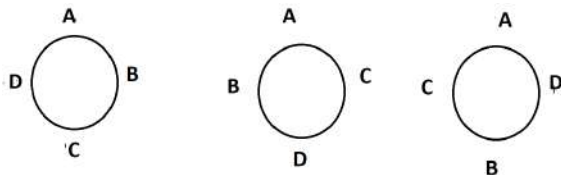
The number of ways n distinct objects can be arranged in a circle (clockwise and anticlockwise are same) is equal to $\frac{(n-1)!}{2}$

Example : Solving for n

▶ Find value of n if ${}^nP_4 = 20\, {}^nP_2$
  Answer: $\frac{n!}{(n-4)!} = 20 \times \frac{n!}{(n-2)!}$
  Solving $(n-2) \times (n-3) = 20$, we get $n = -2$ or $n = 7$.
  Eliminating $n = -2$, we get $n = 7$.

Example : Solving for n

- $\frac{^nP_4}{^{n-1}P_4} = \frac{5}{3}$

  Answer: $\frac{n!}{(n-4)!} \times \frac{(n-5)!}{(n-1)!} = \frac{5}{3}$

  $\frac{n}{(n-4)} = \frac{5}{3}$

  Solving for n gives us $n = 10$.

Example : Solving for r

▶ Find r, if ${}^5P_r = 2.{}^6P_{r-1}$
  Answer: $\frac{5!}{(5-r)!} = 2.\frac{6!}{(7-r)!}$
  $\frac{5!}{(5-r)!} = 2.\frac{6!}{(7-r)(6-r)(5-r)!}$
  Solving $(7-r)(6-r) = 12$ gives $r = 10$ or $r = 3$.

Since $r \leq n$, the option $r = 10$ is eliminated and we get $r = 3$.

Topic summary

1. Permutations when objects are distinct
   1.1 repetitions not allowed.
   1.2 repetitions allowed.

2. Permutations when objects are not distinct.

3. Circular permutations:
   3.1 Clockwise and anticlockwise are different.
   3.2 Clockwise and anticlockwise are same.

4. Solving for r and n using the permutation formula.

Learning objectives

1. Understand basic principles of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.

Combinations

Applications: Permutations or combinations

Introduction

- ▶ Example: How many ways can we select two students from a group of three students?
  - ▶ Let A, B, and C be the three students.
  - ▶ We can chose AB, AC, or BC.
  - ▶ Note, when we talked of permutations, the order was important, i.e., AB was different from BA.
  - ▶ In this case, they are the same- order is not important.
- ▶ Each selection is called a combination of 3 different objects taken 2 at a time.
- ▶ In this case, the concern is only which of the 2 objects are chosen and not in the order in which they are chosen.

Example

Consider A, B, C- Possible combinations- taking two at a time

First place	Second place
A	B
A	C
B	C

- ▶ Note each combination gives rise to 2! arrangements.
- ▶ All combinations give $3 \times 2 = 6$ arrangements.
- ▶ Number of combinations $\times 2! =$ Number of permutations

Combinations: Notation and formula

- In general, each combination of r objects from n objects can give rise to $r!$ arrangements.

- The number of possible combinations of r objects from a collection of n distinct objects is is denoted by ^nC_r and is given by

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

- Another common notation is $\binom{n}{r}$ which is also referred to as the binomial coefficient

Some useful results

1. ${}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{(n-r)!r!} = {}^{n}C_{(n-r)}$
   In other words, selecting r objects from n objects is the same as rejecting $n - r$ objects from n objects.

2. ${}^{n}C_{n} = 1$ and ${}^{n}C_{0} = 1$ for all values of n

3. ${}^{n}C_{r} = {}^{n-1}C_{r-1} + {}^{n-1}C_{r}; 1 \leq r \leq n$

Example: Choosing questions in an exam

- ▶ In an examination, a question paper consists of 12 questions divided into two parts i.e., Part I and Part II, containing 7 and 5 questions, respectively. A student is required to attempt 8 questions in all, selecting at least 3 from each part. In how many ways can a student select the questions ?

- ▶ Solution: $^7C_3\,^5C_5 + \,^7C_4\,^5C_4 + \,^7C_5\,^5C_3 = 35 + 175 + 210 = 420$

Example: Game of cards

Lets consider the case of choosing four cards from a deck of 52 cards.
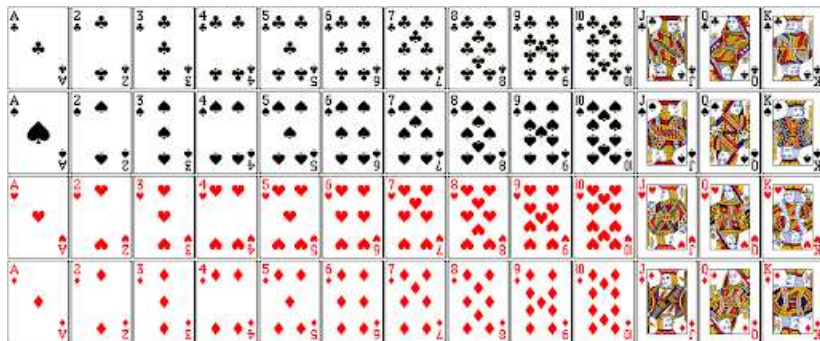
Example: Game of cards contd.

1. Total number of ways of choosing four cards from 52 cards $=$
   $^{52}C_4 = \frac{52!}{4!48!} = 2,70,725$

2. All four cards are of the same suit
   $^4C_1 \times {}^{13}C_4 = 4 \times \frac{13!}{4!9!} = 2860$

3. Cards are of same colour $^2C_1 \times {}^{26}C_4 = 2 \times \frac{26!}{4!22!} = 2,99,00$

Example: Choosing a cricket team

▶ Select a cricket team of eleven from 17 players in which only 5 players can bowl. The requirement is the cricket team of 11 must include exactly 4 bowlers? How many ways can the selection be done?

▶ Solution:

   ▶ Total number of players available for selection: 17
     Number of bowlers: 5
   ▶ Need four bowlers: This selection can be done in ${}^5C_4$ ways.
   ▶ Remaining seven players can be selected from remaining twelve players in ${}^{12}C_7$ ways.
   ▶ Total number of ways the selection can be done is
     ${}^5C_4 \times {}^{12}C_7 = 5 \times 792 = 3960$ ways

Example: Drawing lines in a circle

▶ Given n points on a circle, how many lines can be drawn connecting these points?

▶ $n = 2$ points, one line can be drawn connecting the points

line segment: AB

▶ $n = 3$ points, three line can be drawn connecting the points

line segments: AB, AC, and BC

▶ In general, given n points, number of line segments that can be drawn connecting the points is ${}^{n}C_2$

Section summary

1. Notation and formula for selecting r objects from n objects.
2. Some useful combinatorial identities.

Learning objectives

1. Understand basic principles of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.

Applications: Permutations or combinations

Applications: Permutations or combinations

▶ Important to distinguish between situations involving combinations and situations involving permutations.

▶ Permutation- "order matters". Combination - "order does not matter"

Example: Finishing a race

▶ Consider the situation of eight athletes participating in a
100m race in a competition with several rounds.

  1. How many different ways can you award the Gold, Silver, and
     Bronze medals?
  2. How many different ways can you choose the top three
     athletes to proceed to the next round in the competition?

▶ Solution:

  1. How many different ways can you award the Gold, Silver, and
     Bronze medals?
     Order is important- Hence we need permutation. Answer is
     $^{8}P_{3} = 336$ ways.
  2. How many different ways can you choose the top three
     athletes to proceed to the next round in the competition?
     Order is not important- Hence we need combination. Answer
     is $^{8}C_{3} = 56$ ways.

Example: Selecting a team

▶ Consider the situation of a class with forty students.
   1. How many different ways can we choose two leaders?
   2. How many different ways can we choose a captain and vice captain?

▶ Solution:
   1. How many different ways can we choose two leaders? Order not important- -hence, combination Answer: $^{40}C_2 = 780$ ways
   2. How many different ways can we choose a captain and vice captain? Order important- -hence, permutation Answer: $^{40}P_2 = 1560$ ways

Example: Drawing lines in a circle

▶ Given *n* points on a circle, how many lines can be drawn connecting these points?

▶ Solution:

1. If the segment has a direction line segment AB is different from BA. Order is important. Hence, total number of ways is $^nP_2$
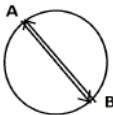
2. If segment has no direction. Line segment AB. Order is not important. Hence, total number of ways is $^nC_2$.
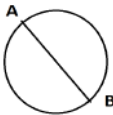
Section summary

- ▶ Need to distinguish between permutation and combination.
- ▶ Examples of situations where permutation is applied, combination is applied.

Learning objectives

1. Understand uncertainty and concept of a random experiment.
2. Describe sample spaces, events of random experiments.
3. Understand the notion of simple event and compound events.
4. Basic laws of probability.
5. Calculate probabilities of events and use a tree diagram to compute probabilities.
6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
7. Distinguish between independent and dependent events.
8. Solve applications of probability.

Random Experiment, Sample Space, Events

Introduction

- ▶ There is a 50% chance that India will win the toss.
- ▶ My guess is answer "a" is the right choice.
- ▶ Party ABC will probably win the next election.
- ▶ There is a 30% chance of rain tomorrow.

- ▶ We routinely see or hear claims as the ones mentioned above. What do they mean?
- ▶ Indeed, as a general rule, to be able to draw valid inferences about a population from a sample, one needs to know how likely it is that certain events will occur under various circumstances.
- ▶ The determination of the likelihood, or chance, that an event will occur is the subject matter of **probability**.

Random experiment

Definition
An experiment is any process that produces an observation or outcome.

Definition
A random experiment is an experiment whose outcome is not predictable with certainty.

Remark
However, although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known.

Examples of random experiments

▶ Experiment: Guessing answers to a four option multiple choice question:
Outcome: A,B,C,D

▶ Experiment: Order of finish in a race with six students- A, B, C, D, E, F.
Outcome: all possible permutations of A, B, C, D, E,and F.

▶ Experiment: Tossing two coins and noting the outcomes
Outcome: HH, HT, TH, TT

▶ Experiment: Measuring the lifetime (in hours) of a bulb
Outcome: 0, or 1 hour, or 2 hours, or,....so on.

▶ Experiment: To throw a dart on a unit square and note the point where it lands.
Outcome: Any point in the square (assuming the dart lands within the square).

Sample Space

Definition

A sample space (denoted by Ω or S) : collection of all basic outcomes.

- ▶ Basic Outcomes: the possible outcomes that can occur must be:
    1. mutually exclusive: only one basic outcome can occur
    2. exhaustive: one basic outcome must occur

Examples of sample spaces

▶ Experiment: Guessing answers to a four option multiple choice question:
Sample space: $S = \{A, B, C, D\}$

▶ Experiment: Order of finish in a race with six students-A, B, C, D, E, F.
Sample space: $S = \{ABCDEF, ABCDFE, \ldots, EFDBAC\}$

▶ Experiment: Tossing two coins and noting the outcomes
Sample space: $S = \{HH, HT, TH, TT\}$

▶ Experiment: Measuring the lifetime (in hours) of a bulb
Sample space: $S = \{x : 0 \leq x < \infty\}$

▶ Experiment: To throw a dart on a unit square and note the point where it lands.
Sample space: $S = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$

Section summary

.

- ▶ Random experiment
- ▶ Sample space: set of all basic outcomes of a random experiment.

Learning objectives

1. Understand uncertainty and concept of a random experiment.
2. Describe sample spaces, events of random experiments.
3. Understand the notion of simple event and compound events.
4. Basic laws of probability.
5. Calculate probabilities of events and use a tree diagram to compute probabilities.
6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
7. Distinguish between independent and dependent events.
8. Solve applications of probability.

Random Experiment, Sample Space, Events

Events

Definition

An event E is a collection of basic outcomes.

- ▶ That is, an event is a subset of the sample space.
- ▶ We say an event has occurred if the outcome is contained in the subset.

Examples of events

▶ Experiment: Guessing answers to a four option multiple choice question:
Event: answer is A; $E = \{A\}$

▶ Experiment: Order of finish in a race with six students- A, B, C, D, E, F.
Event: A finishes the race first
$E = \{ABCDEF, ABCDFE, ABDCFE, \ldots, AFEDBC\}$

▶ Experiment: Tossing two coins and noting the outcomes
Event: head on the first toss $E = \{HH, HT\}$

▶ Experiment: Measuring the lifetime (in hours) of a bulb
Event: life time is less than or equal to four hours
$E = \{x : 0 \leq x \leq 4\}$

Union of events

- ▶ For any two events E and F, we define the new event $E \cup F$ called the union of events E and F, to consist of all outcomes that are in E or in F or in both E and F.
- ▶ That is, the event $E \cup F$ will occur if either E or F occurs.

Examples of union of events

- ▶ Experiment: Guessing answers to a four option multiple choice question:
  Event:
  - ▶ answer is A; $E_1 = \{A\}$
  - ▶ answer is B; $E_2 = \{B\}$
  - ▶ answer is A or B; $E_3 = E_1 \cup E_2 = \{A, B\}$
- ▶ Experiment: Order of finish in a race with six students- A, B, C, D, E, F.
  Event:
  - ▶ A finishes the race first
    $E_1 = \{ABCDEF, ABCDFE, ABDCFE, \ldots, AFEDBC\}$
  - ▶ B comes second in the race
    $E_2 = \{ABCDEF, ABCDFE, ABDCFE, \ldots, CBADEF\}$
  - ▶ A comes first **or** B comes second.
    $E_1 \cup E_2 =$
    $\{\textbf{ABCDEF}, \textbf{ABCDFE}, \textbf{ABDCFE}, \ldots, AFEDBC, CBADEF\}$

Examples

- ▶ Experiment: Tossing two coins and noting the outcomes
  Event:
    - ▶ head on the first toss $E_1 = \{HH, HT\}$
    - ▶ head on second toss $E_2 = \{HH, TH\}$
    - ▶ head on first or second toss $E_1 \cup E_2 = \{HH, HT, TH\}$

Intersection of events

- For any two events E and F, we define the new event $E \cap F$ called the intersection of events E and F, to consist of all outcomes that are in E and in F.

- That is, the event $E \cap F$ will occur if both E and F occurs.

Examples

- ▶ Experiment: Order of finish in a race with six students-
  A, B, C, D, E, F.
  Event:
  - ▶ A finishes the race first
    $E_1 = \{ABCDEF, ABCDFE, ABDCFE, \ldots, AFEDBC\}$
  - ▶ B comes second in the race
    $E_2 = \{ABCDEF, ABCDFE, ABDCFE, \ldots, CBADEF\}$
  - ▶ A comes first **and** B comes second.
    $E_1 \cap E_2 = \{\textbf{ABCDEF}, \textbf{ABCDFE}, \textbf{ABDCFE}, \ldots, \textbf{ABDCFE}\}$

- ▶ Experiment: Tossing two coins and noting the outcomes
  Event:
  - ▶ head on the first toss $E_1 = \{HH, HT\}$
  - ▶ head on second toss $E_2 = \{HH, TH\}$
  - ▶ head on first and second toss $E_1 \cap E_2 = \{HH\}$

Null event and disjoint event

Definition
We call the event without any outcomes the null event, and designate it as Φ

Definition
If the intersection of E and F is the null event, then since E and F cannot simultaneously occur, we say that E and F are disjoint, or mutually exclusive.

Examples of null event

▶ Experiment: Guessing answers to a four option multiple choice question:

  Event:

  ▶ answer is A; $E_1 = \{A\}$
  ▶ answer is B; $E_2 = \{B\}$
  ▶ answer is A and B; $E_3 = E_1 \cap E_2 = \Phi$
  ▶ We say events E_1 and E_2 are mutually exclusive or disjoint. Occurence of E_1 disallows occurence of E_2. In other words, if my $A(B)$ is my guess, then $B(A)$ cannot be my guess.

Complement of an event

Definition

The complement of E, denoted by E^c, consists of all outcomes in the sample space S that are not in E.

- ▶ That is, E^c will occur if and only if E does not occur.
- ▶ The complement of the sample space is the null set, that is $S^c = \Phi$

Examples of complement of an event

▶ Experiment: Toss a coin once and note the outcomes
  ▶ Sample space: $S = \{H, T\}$
  ▶ Event E_1: out come is head $E_1 = \{H\}$
  ▶ Event E_2: out come is tail $E_2 = \{T\}$
  ▶ Event E_2 is complement of event E_1. In other words, $E_2 = E_1^c$

▶ Experiment: Tossing two coins and noting the outcomes
  ▶ Sample space: $S = \{HH, HT, TH, TT\}$
  ▶ Event: head on the first toss $E_1 = \{HH, HT\}$
  ▶ $E_1^c = \{TH, TT\}$; tail on first toss

Subsets

Definition

For any two events E and F, if all of the outcomes in E are also in F, then we say that E is contained in F, or E is a subset of F, and denote it as $E \subset F$

- ▶ Example: Experiment: Tossing two coins and noting the outcomes
  - ▶ Sample space: $S = \{HH, HT, TH, TT\}$
  - ▶ Event: head on the first toss $F = \{HH, HT\}$
  - ▶ Event: head in both the tosses $E = \{HH\}$
  - ▶ $E \subset F$

Section summary

1. Notion of events
2. Union, intersection, complement of events
3. Null event and mutually exclusive (disjoint) events

Learning objectives

1. Understand uncertainty and concept of a random experiment.
2. Describe sample spaces, events of random experiments.
3. Understand the notion of simple event and compound events.
4. Basic laws of probability.
5. Calculate probabilities of events and use a tree diagram to compute probabilities.
6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
7. Distinguish between independent and dependent events.
8. Solve applications of probability.

Venn diagrams

Application: playing cards

▶ A deck of playing cards is a collection of 52 playing cards
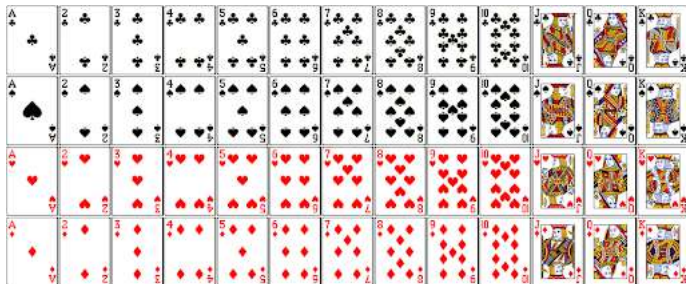
▶ Experiment: Randomly selecting one card from the deck, we will get one of these 52 cards

▶ Sample space: $S = \{$collection of all 52 cards$\}$

Application: playing cards contd.

▶ Describe the event that the card selected is the king of hearts.

▶ Describe the event that the card selected is a king.

▶ Describe the event that the card selected is hearts.

Application: playing cards contd.

▶ Determine and describe the event $F \cup G$
  Event $F \cup G$ implies the card selected is a King or a heart.
  Hence the outcome of $F \cup G$ is
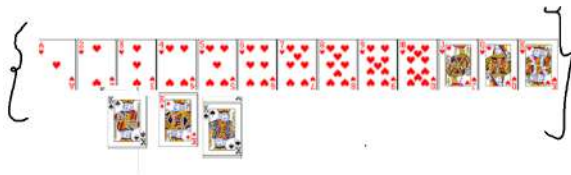
Application: playing cards contd.

▶ Determine and describe the event $F \cap G$
Event $F \cap G$ implies the card selected is a King and a heart.
Hence the outcome of $F \cap G$ is

which is same as event E

Application: playing cards contd.

▶ Let H be the event of selecting an Ace. Are events G and H mutually exclusive?

▶ Event G and event H are not mutually exclusive because they have the common outcome "ace of hearts." Both events occur if the card selected is the ace of hearts.

▶ Let I be the event of selecting a Queen. Are events F and I mutually exclusive? Yes. If we select a king card, we cannot select a queen card.

Venn diagrams

- A graphical representation that is useful for illustrating logical relations among events is the Venn diagram.

Representation of event

▶ Representation of event: The events $E, F, G, \ldots$ are represented in given circles within the rectangle.
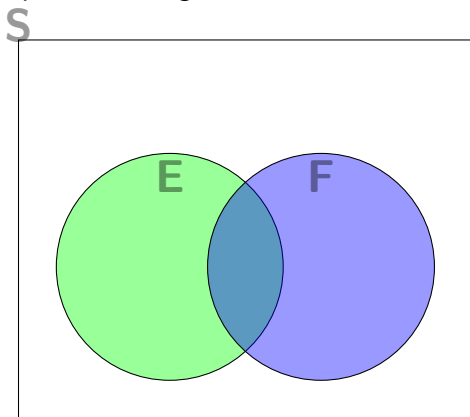
Representation of event: union and intersection

- ▶ Representation of event:
  - ▶ $E \cup F$ is entire shaded region
  - ▶ $E \cap F$ is the shaded in blue region
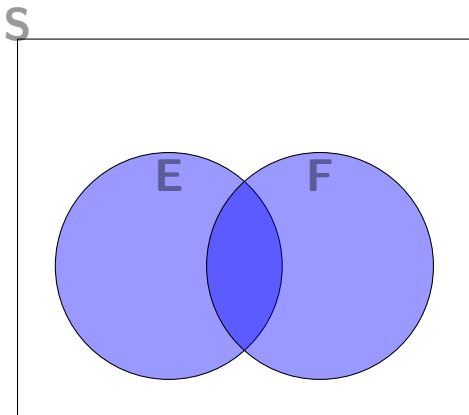
Representation of event: disjoint events
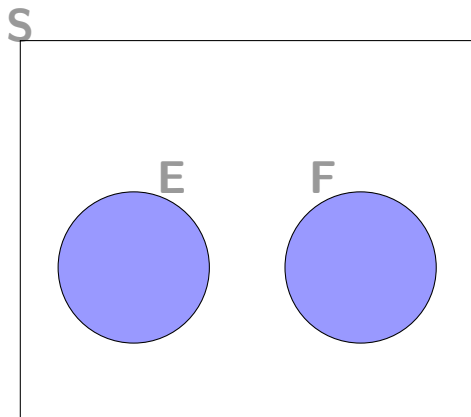
Representation of event: subsets
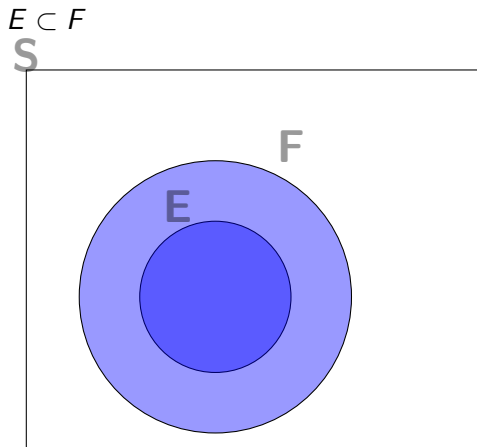
Topic summary

1. Introduced random experiment, sample space, event.
2. Notion of union, intersection, complement of events.
3. Representation of sample space, events, using venn diagrams.

Learning objectives

1. Understand uncertainty and concept of a random experiment.
2. Describe sample spaces, events of random experiments.
3. Understand the notion of simple event and compound events.
4. Basic laws of probability.
5. Calculate probabilities of events and use a tree diagram to compute probabilities.
6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
7. Distinguish between independent and dependent events.
8. Solve applications of probability.

Properties of Probability
  Equally likely outcomes

The three main interpretations of probability

1. Classical (Apriori or theoretical): Let S be the sample space of a random experiment in which there are n equally likely outcomes, and the event E consists of exactly m of these outcomes, then we say the probability of the event E is $\frac{m}{n}$ and represent it as $P(E) = \frac{m}{n}$

2. Relative frequency (Aposteriori or empirical): The probability of an event in an experiment is the proportion (or fraction) of times the event occurs in a very long (theoretically infinite) series of (independent) repetitions of experiment. In other words, if $n(E)$ is the number of times E occurs in n repetitions of the experiment, $P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$

3. Subjective: The probability of an event is a "best guess" by a person making the statement of the chances that the event will happen. The probability measures an individual's degree of belief in the event.

Probability Axioms

Consider an experiment whose sample space is S. We suppose that for each event E there is a number, denoted $P(E)$ and called the probability of event E, that is in accord with the following three properties (axioms).

1. For any event E, the probability of E is a number between 0 and 1. That is, $0 \leq P(E) \leq 1$.

2. The probability of sample space S is 1. Symbolically, $P(S) = 1$. In other words, the outcome of the experiment will be an element of sample space S with probability 1.

3. For a sequence of mutually exclusive (disjoint) events, $E_1, E_2, \ldots,$

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

Probability of union of disjoint events

The third property can be stated as:
The probability of the union of disjoint events is equal to the sum of the probabilities of these events.
For instance, if E_1 and E_2 are disjoint, then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

In other words, if events E_1 and E_2 cannot simultaneously occur, then the probability that the outcome of the experiment is contained in either E_1 or E_2 is equal to the sum of the probability that it is in E_1 and the probability that it is in E_2.

General properties of probability

Properties 1, 2, and 3 can be used to establish some general results concerning probabilities.

1. Probability of complement of an event: $P(E^c) = 1 - P(E)$
    - E and E^c are disjoint. Also, $E \cup E^c = S$
    - Apply Property 3 to LHS $P(E \cup E^c) = P(E) + P(E^c)$
    - Apply Property 2 to RHS $P(S) = 1$
    - Equating both, we get
      $P(E \cup E^c) = P(E) + P(E^c) = P(S) = 1$.
      Hence $P(E^c) = 1 - P(E)$

2. $P(\Phi) = 0$
    - $S^c = \Phi$
    - Apply the above property, $P(S^c) = 1 - P(S)$
    - Apply property 2, $P(S) = 1$
    - Hence $P(\Phi) = 0$

Addition rule of probability

The following formula relates the probability of the union of events
E_1 and E_2, which are not necessarily disjoint, to $P(E_1), P(E_2)$, and
the probability of the intersection of E_1 and E_2. It is often called
the addition rule of probability.

For any events E_1 and E_2,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Proof of addition rule

- $E_1 \cup E_2 = I \cup II \cup III$
- $E_1 = I \cup II$
- $E_2 = II \cup III$
- $E_1 \cap E_2 = II$
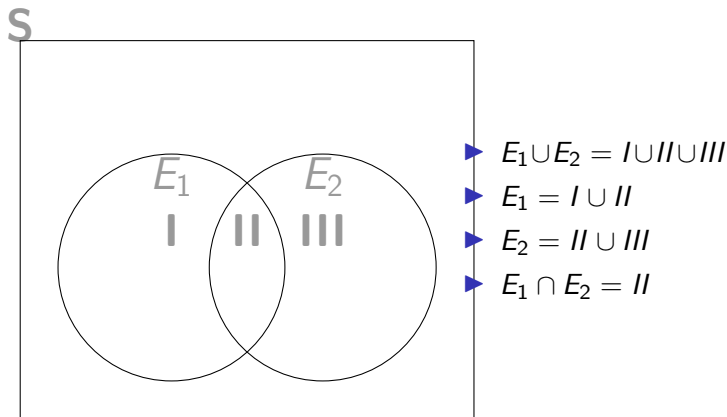
Section summary

▶ Probability axioms
▶ Properties of Probability
  ▶ Addition rule

Learning objectives

1. Understand uncertainty and concept of a random experiment.
2. Describe sample spaces, events of random experiments.
3. Understand the notion of simple event and compound events.
4. Basic laws of probability.
5. Calculate probabilities of events and use a tree diagram to compute probabilities.
6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
7. Distinguish between independent and dependent events.
8. Solve applications of probability.

Application of addition rule

Example: Shopping for shirts and pants

A customer that goes to the clothing store will purchase a shirt with probability 0.3. The customer will purchase a pant with probability 0.2 and will purchase both a shirt and a pant with probability 0.1. What proportion of customers purchases neither a shirt nor a pant?

- ▶ Let S denote the event of a customer purchasing a shirt
- ▶ Let P denote the event of a customer purchasing a pant
- ▶ Proportion of customers purchases neither a shirt nor a pant?
- ▶ Neither a shirt nor a pant is complement of the event of either shirt or pant. What we seek is $P(S \cup P)^c$
- ▶ We know $P(S \cup P)^c = 1 - P(S \cup P)$
- ▶ $P(S \cup P) = P(S) + P(P) - P(S \cap P) = 0.3 + 0.2 - 0.1 = 0.4$
- ▶ Hence, $P(S \cup P)^c = 1 - 0.4 = 0.6$

Example: subject grades

A student has a 40 percent chance of receiving an A grade in statistics, a 60 percent chance of receiving an A in mathematics, and an 86 percent chance of receiving an A in either statistics or mathematics. Find the probability that she

1. Does not receive an A in either statistics or mathematics.

2. Receives A's in both statistics and mathematics.

▶ Let S denote the event of obtaining a A grade in statistics

▶ Let M denote the event of obtaining a A grade in mathematics

▶ Event does not receive an A in either statistics or mathematics= complement of the event that student receives an A in at least one of the subjects= $(S \cup M)^c$.

▶ Event receives A's in both statistics and mathematics= $(S \cap M)$

Example: subject grades-contd.

- $P(S \cup M)^c = 1 - P(S \cup M) = 1 - .86 = 0.14$
- $P(S \cap M) = P(S) + P(M) - P(S \cup M) = 0.40 + 0.60 - 0.86 = 0.14$

Section summary

1. Addition rule of probability.

Learning objectives

1. Understand uncertainty and concept of a random experiment.
2. Describe sample spaces, events of random experiments.
3. Understand the notion of simple event and compound events.
4. Basic laws of probability.
5. Calculate probabilities of events and use a tree diagram to compute probabilities.
6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
7. Distinguish between independent and dependent events.
8. Solve applications of probability.

Equally likely outcomes

Equally likely outcomes

▶ For certain experiments it is natural to assume that each outcome in the sample space S is equally likely to occur.

▶ That is, if sample space S consists of N outcomes, say, $S = \{1, 2, \ldots, N\}$, then it is often reasonable to suppose that

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{N\})$$

▶ In this expression, $P(\{i\})$ is the probability of the event consisting of the single outcome i.

▶ Using the properties of probability, we can show that the foregoing implies that the probability of any event A is equal to the proportion of the outcomes in the sample space that is in A.

▶ That is, $P(A) = \dfrac{\text{number of outcomes in } S \text{ that are in } A}{N}$

Example: Rolling a dice

▶ Experiment: Roll a fair dice

▶ Sample space: $S = \{1, 2, 3, 4, 5, 6\}$

▶ Let E_i denote the event of outcome i. Since the dice is fair, $P(E_i) = \frac{1}{6}$.

▶ Define A to be the event the outcome is odd $A = \{1, 3, 5\}$
$P(A) = P(E_1) + P(E_3) + P(E_5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

▶ Let B be the event that the outcome is greater than 4.
$B = \{5, 6\}$ $P(B) = \frac{2}{6}$

▶ Let C be the event that the outcome is either odd or greater than 4.
$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}$

Example: playing cards

When drawing a card from a standard deck of 52 playing cards, what is the probability that the card is either red or a queen?
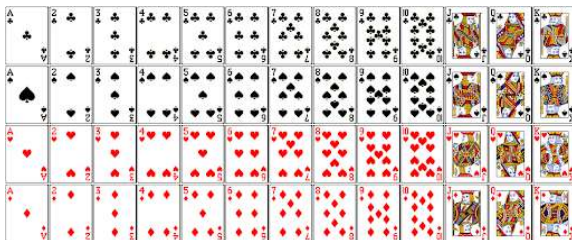
▶ Let R be the event that the card drawn is Red.
$P(R) = \frac{26}{52} = \frac{1}{2}$

▶ Let Q be the event that the card drawn is Queen.
$P(Q) = \frac{4}{52} = \frac{1}{13}$

Example: playing cards-contd.

Probability that the card is either red or a queen= $P(R \cup Q)$

▶ Applying addition rule: $P(R \cup Q) = P(R) + P(Q) - P(R \cap Q)$

▶ $R \cap Q$ describes the event that the card drawn in a Red Queen. $P(R \cap Q) = \frac{2}{56}$

▶ Hence $P(R \cup Q) = \frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13}$

Topic summary

1. Interpretations of probability
2. Probability axioms
3. Addition rule of probability.
4. Equally likely outcomes.

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Contingency tables: Joint, Marginal, and Conditional probabilities

From tables to probability

- ▶ Recall the cell phone usage versus gender example when we discussed about association between categorical variables and the concept of relative frequencies.

- ▶ Percentages computed within rows or columns of a contingency table correspond to conditional probabilities

- ▶ Convert contingency tables into probabilities, we use the counts to define probabilities.

Relative frequency

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10	34	**44**
Male	14	42	**56**
Column total	**24**	**76**	**100**

Divide each count by 100

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10/100	34/100	**44/100**
Male	14/100	42/100	**56/100**
Column total	**24/100**	**76/100**	**100**

Joint probabilities

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	0.10	0.34	**0.44**
Male	0.14	0.42	**0.56**
Column total	**0.24**	**0.76**	**100**

▶ Displayed in cells of a contingency table

▶ Represent the probability of an intersection of two or more events

▶ In the example: there are four joint probabilities; e.g.,
  ▶ P(Female **and** Not owning a smartphone) $= 0.10$
  ▶ P(Male and Owning a smartphone) $= 0.42$

Marginal probability

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	0.10	0.34	**0.44**
Male	0.14	0.42	**0.56**
Column total	**0.24**	**0.76**	**100**

▶ Displayed in the margins of a contingency table

▶ Is the probability of observing an outcome with a single attribute, regardless of its other attributes

▶ In the example: There are four marginal probabilities, e.g.,
  ▶ P(Female) $= 0.10 + 0.34 = 0.44$
  ▶ P(Owning a smartphone) $= 0.34 + 0.42 = 0.76$

Conditional probability

▶ Find conditional probabilities to answer questions like
  ▶ "among Female buyers, what is the chance a someone owns a phone?"
  ▶ "Among people who don't own a phone, how many are male?"
▶ Recognize the answers
  ▶ "among Female buyers, what is the chance a someone owns a phone?"- row relative frequency
  ▶ "Among people who don't own a phone, how many are male?"- column relative frequency

Conditional probability

We restrict the sample space to a row or column.

▶ "among Female buyers, what is the chance a someone owns a phone?"- Restrict sample space to only "Females"- First row

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10/44	34/44	**44**
Male	14/56	42/56	**56**
Column total	**24/100**	**76/100**	**100**

$P(\text{Doesn't own a phone}|\text{Female}) = \frac{10}{44} =$
$\frac{P(\text{Female} \cap \text{Doesn't own a phone})}{P(\text{Female})}$

Conditional probability

We restrict the sample space to a row or column.

▶ "Among people who don't own a phone, how many are male?"- Restricting sample space to only people who "don't own a phone"- First column

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
Female	10/24	34/76	**44/100**
Male	14/24	42/76	**56/100**
Column total	**24**	**76**	**100**

$P(\text{Female}|\text{Doesn't own a phone}) = \frac{10}{24} = \frac{P(\text{Female} \cap \text{Doesn't own a phone})}{P(\text{Doesn't own a phone})}$

Section summary

▶ Revisited contingency tables and introduced notions of
   1. Joint probability
   2. Marginal probability
   3. Conditional probability

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Conditional Probability

Multiplication rule

Independent events

Bayes' rule

Introduction

▶ We are often interested in determining probabilities when some partial information concerning the outcome of the experiment is available. In such situations, the probabilities are called conditional probabilities.

Example: Roll a dice twice

- ▶ Experiment: Roll a dice twice
- ▶ Sample space:
$$S = \left\{ \begin{array}{l} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6), \end{array} \right\}$$
- ▶ Each outcome is equally likely to occur with a probability of $\frac{1}{36}$

Example: Rolling a dice twice- contd.

▶ Suppose further that the first roll of the dice lands on 4.

▶ Given this information, what is the resulting probability that the sum of the dice is 10?

▶ In other words, the restricted sample space if the first dice lands of a four $F = \{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\}$

▶ If each outcome of a finite sample space S is equally likely, then, conditional on the event that the outcome lies in a subset F, all outcomes in F become equally likely. In such cases, it is often convenient to compute conditional probabilities of the form $P(E|F)$ by using F as the sample space.

▶ Among outcomes in the restricted sample space, the outcome that satisfies the sum of dice is 10 is outcome $(4, 6)$. And this happens with Probability $\frac{1}{6}$

Conditional Probability: formula

▶ Let E denote the event that the sum of the dice is 10 and let F denote the event that the first die lands on 4, then the probability obtained is called the conditional probability of E given that F has occurred. It is denoted by

$$P(E|F)$$

▶ The probability that event E occurs given that event F occurs (or conditional on event F occurring) is given by

$$P(E|F) = \frac{P(E \cap F)}{P(F)}; \ P(F) > 0$$

Conditional probability: Venn diagram illustration
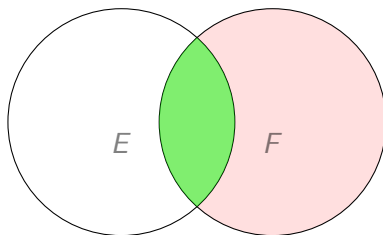
$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Apply the formula to the example

▶ As a further check of the preceding formula for the conditional probability, use it to compute the conditional probability that the sum of a pair of rolled dice is 10, given that the first die lands on 4.

▶ $P(E|F) = \frac{P(E \cap F)}{P(F)}$

▶ $\frac{P(E \cap F)}{P(F)} = \frac{P(\{(4,6)\})}{P(\{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\})} = \frac{1/36}{6/36} = \frac{1}{6}$

Section summary

1. Introduced notion of conditional probability
2. Formula: $P(E|F) = \frac{P(E \cap F)}{P(F)}$

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Multiplication rule

Multiplication rule

▶ Conditional probability formula:

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

Multiply both sides of the formula with $P(F)$, we get

▶

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

▶ This rule states that the probability that both E and F occur is equal to the probability that F occurs multiplied by the conditional probability of E given that F occurs.

▶ It is often quite useful for computing the probability of an intersection.

Example: application of multiplication rule

▶ In an introductory statistics class of forty students, the number of males is equal to 23 and number of females is equal to 17. Two students are selected at random from the class. The first student selected is not returned to the class for possible reselection; that is, the sampling is without replacement. Find the probability that the first student selected is female and the second is male.

▶ Experiment: Selecting two students from forty students.

▶ Sample space: $S = \{M_1 M_2, M_1 F_2, F_1 M_2, F_1 F_2\}$; where $M_1 M_2$ represents the outcome the first student is male and the second student is male. Other outcomes can be interpreted similarly.

Example: application of multiplication rule-cont.

- ▶ Event: First student female and second is male.
- ▶ $P(F_1 M_2) =$
  $P(\text{First student is female} \cap \text{Second student is male})$
- ▶ $P(\text{First student is female}) = \frac{17}{40}$.
- ▶ Given sampling without replacement; first student selected is not returned to the class for reselection.
- ▶ Given that the first student selected is female, of the 39 students remaining in the class 23 are male, so
  $P(\text{Second student is male}|\text{First student is Female}) = \frac{23}{39}$
- ▶ Hence, $P(\text{First student female and second is male}) =$
  $P(\text{First student is female}) \times$
  $P(\text{Second student is male}|\text{First student is Female}) = \frac{17}{40}\frac{23}{39} = 0.251$

Generalized multiplication rule

A generalization of the multiplication rule, which provides an expression for the probability of the intersection of an arbitrary number of events, is referred to as the generalized multiplication rule and is given by $P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)\ldots P(E_n|E_1 \cap E_2 \ldots \cap E_{n-1})$

Example: deck of cards

▶ An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. Compute the probability that each pile has exactly 1 ace.

▶ Define events E_i; $i = 1, 2, 3, 4$ as follows
  1. $E_1 = \{$the ace of spades is in any one of the piles$\}$
  2. $E_2 =$
     $\{$the ace of spades and the ace of hearts are in different pile$\}$
  3. $E_3 =$
     $\{$the aces of spades, heart, and diamonds are in different piles$\}$
  4. $E_4 = \{$all four aces are in different piles$\}$

▶ What we require is $P(E_1 \cap E_2 \cap E_3 \cap E_4)$

Solution

1. $P(E_1) = 1$
2. $P(E_2|E_1) = \frac{39}{51}$
3. $P(E_3|E_1 \cap E_2) = \frac{26}{50}$
4. $P(E_4|E_1 \cap E_2 \cap E_3) = \frac{13}{49}$
5. $P(E_1 \cap E_2 \cap E_3 \cap E_4) = \frac{39}{51} \frac{26}{50} \frac{13}{49} \approx 0.105$

Section summary

1. Multiplication rule and its application to find probability of intersection of events.

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Independent events

Independent events

▶ Question: Will the conditional probability that E occurs given that F has occurred be generally equal to the (unconditional) probability of E?

▶ That is, Will knowing that F has occurred generally change the chances of E's occurrence?

▶ In the cases where $P(E|F)$ is equal to $P(E)$, we say that E is independent of F.

  ▶ In other words, event E is independent of event F if knowing whether F occurs does not affect the probability of E.

Independent events: definition

Since
$$P(E \cap F) = P(F) \times P(E|F)$$

we see that E is independent of F if

$$P(E \cap F) = P(F) \times P(E)$$

Definition
Two events E and F are independent if $P(E \cap F) = P(E) \times P(F)$.

Definition
Two events that are not independent are said to be dependent.

Multiplication rule for two independent events

▶ For any two events, E and F, If E and F are independent events, then

$$P(E \cap F) = P(E) \times P(F)$$

and conversely, if

$$P(E \cap F) = P(E) \times P(F)$$

then E and F are independent.

▶ In other words, two events are independent if and only if the probability that both occur equals the product of their individual probabilities.

▶ The definition of independence for three or more events is more complicated than that for two events. We will discuss this later.

Section summary

- ▶ Independent events
- ▶ Multiplication rule for two independent events.

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Independent events: example
    Rolling a dice
    Deck of cards

Example: Roll a dice twice

▶ Experiment: Roll a dice twice

▶ Sample space:

$$S = \left\{ \begin{array}{l} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6), \end{array} \right\}$$

▶ Define the following events

    ▶ E_1: The first outcome is a 3

    ▶ E_2: Sum of outcomes is 8

    ▶ E_3: Sum of outcomes is 7

▶ Are events E_1 and E_2 independent?

▶ Are events E_1 and E_3 independent?

Statistics for Data Science -1
└─ Independent events: example
  └─ Rolling a dice

Are E_1 and E_2 independent?-solution

▶ $E_1 \cap E_2$ is the event that the first outcome is 3 and sum of outcomes is 8.

$$P(E_1 \cap E_2) = P(\{(3,5)\}) = \frac{1}{36}$$

▶ $P(E_1) = P(\{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}) = \frac{6}{36}$
▶ $P(E_2) = P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36}$
▶ Since $\frac{1}{36} \neq \frac{6}{36} \times \frac{5}{36}$ we see that $P(E_1 \cap E_2) \neq P(E_1) \times P(E_2)$, so events E_1 and E_2 are not independent.

Are E_1 and E_3 independent?-solution

- $E_1 \cap E_3$ is the event that the first outcome is 3 and sum of outcomes is 7.

$$P(E_1 \cap E_3) = P(\{(3,4)\}) = \frac{1}{36}$$

- $P(E_1) = P(\{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}) = \frac{6}{36}$
- $P(E_3) = P(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = \frac{6}{36}$
- Since $\frac{1}{36} = \frac{6}{36} \times \frac{6}{36}$ we see that $P(E_1 \cap E_3) = P(E_1) \times P(E_3)$, so events E_1 and E_3 are independent.

Example: deck of cards

Consider again the experiment of randomly selecting one card from a deck of 52 playing cards.

- ▶ Define the following events
    - ▶ E_1: A face card is selected.
    - ▶ E_2: A king is selected.
    - ▶ E_3: A heart is selected.
- ▶ Are E_1 and E_2 independent?
- ▶ Are E_2 and E_3 independent?

Are E_1 and E_2 independent?-solution

▶ $E_1 \cap E_2$ is the event that a face card and a king is selected which is the event a king is selected.

$$P(E_1 \cap E_2) = P(\{KH, KC, KS, KD\}) = \frac{4}{52}$$

▶ $P(E_1) =$
  $P(\{JH, JC, JS, JD, KH, KC, KS, KD, QH, QC, QS, QD\}) =$
  $\frac{12}{52}$

▶ $P(E_2) = P(\{KH, KC, KS, KD\}) = \frac{4}{52}$

▶ Since $\frac{4}{52} \neq \frac{12}{52} \times \frac{4}{52}$ we see that $P(E_1 \cap E_2) \neq P(E_1) \times P(E_2)$, so events E_1 and E_2 are not independent.

Are E_2 and E_3 independent?-solution

▶ $E_2 \cap E_3$ is the event that a king and a heart is selected which is the event a kingheart is selected.

$$P(E_1 \cap E_2) = P(\{KH\}) = \frac{1}{52}$$

▶ $P(E_2) = P(\{KH, KC, KS, KD\}) = \frac{4}{52}$

▶ $P(E_3) =$
$P(\{AH, 2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, KH, QH\}) = \frac{13}{52}$

▶ Since $\frac{1}{52} = \frac{4}{52} \times \frac{13}{52}$ we see that $P(E_2 \cap E_3) = P(E_2) \times P(E_3)$, so events E_2 and E_3 are independent.

Statistics for Data Science -1
└─ Independent events: example
  └─ Deck of cards

Section summary

▶ Examples of independent and dependent events.

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Independence of E and F^c

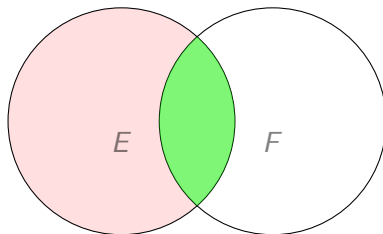Independence of three events

Independence of E and F^c

Proposition

If E and F are independent, then so are E and F^c.

Proof.

▶ Assume E and F are independent.

▶ E can be expressed as

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

Proof continued

▶

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

$E \cap F$ and $E \cap F^c$ are mutually exclusive hence

$$P(E) = P(E \cap F) + P(E \cap F^c)$$

▶ E and F are independent $P(E \cap F) = P(E) \times P(F)$
▶ We get $P(E) = P(E) \times P(F) + P(E \cap F^c)$
▶ Which is equal to $P(E)(1 - P(F)) = P(E \cap F^c)$
▶ Hence, $P(E \cap F^c) = P(E) \times P(F^c)$.

Thus, if E is independent of F, then the probability of E's occurrence is unchanged by information as to whether or not F has occurred.

Independence of more than two events

- Question: Suppose that E is independent of F and is also independent of G. Is E then necessarily independent of $(F \cap G)$?

- Let's go back to the example where two fair dice are thrown. Recall, getting a sum of 7 was independent of the outcome of first throw. Similarly, getting a sum of 7 is independent of the second outcome as well.
  - Let E denote the event that the sum of the dice is 7.
  - Let F denote the event that the first die equals 4
  - Let G denote the event that the second die equals 3.

- $F \cap G$ is the event of first throw is a 4 and second throw is a 3. Now $P(Sum = 7|\text{first throw is 4 and second throw is 3}) = 1$, i.e. $P(E|F \cap G) = 1$. That is, event E is not independent of $(F \cap G)$

Independence of three events

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

1. $P(E \cap F \cap G) = P(E) \times P(F) \times P(G)$
2. $P(E \cap F) = P(E) \times P(F)$
3. $P(E \cap G) = P(E) \times P(G)$
4. $P(F \cap G) = P(F) \times P(G)$

For independent events, the probability that they all occur equals the product of their individual probabilities.

Example: application

▶ A couple is planning on having three children. Assuming that each child is equally likely to be of either sex and that the sexes of the children are independent, find the probability that all three children are girls.

▶ Solution: Define E_i to be the event that the ith child is a girl. The event all three children are girls is $(E_1 \cap E_2 \cap E_3)$

    ▶ Given each child is equally likely to be of either sex $\implies$ $P(E_i) = \frac{1}{2}$

    ▶ the sexes of the children are independent $\implies$ $P(E_1 \cap E_2 \cap E_3) = P(E_1) \times P(E_2) \times P(E_3)$

    ▶ Hence, the probability all three children are girls $=$ $P(E_1 \cap E_2 \cap E_3) = P(E_1) \times P(E_2) \times P(E_3) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$

Section summary

1. Notion of independent events.
   - ▶ Independence of E and F^c.
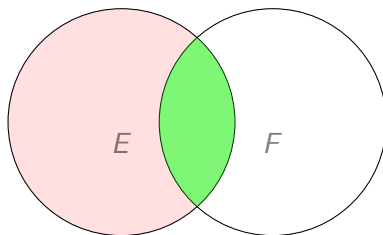2. Independence of more than three events.

Learning objectives

1. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
2. Distinguish between independent and dependent events.
3. Solve applications of probability.

Law of total probability

Bayes' rule

Law of total probability

- Let E and F be events.

- E can be expressed as $(E \cap F) \cup (E \cap F^c)$
- In other words, for, in order for an outcome to be in E, it must either be in both E and F or be in E but not in F.

Formula and interpretation

$$\begin{aligned}
P(E) &= P(E \cap F) + P(E \cap F^c) \\
&= P(F)P(E|F) + P(F^c)P(E|F^c) \quad\quad (1)
\end{aligned}$$

▶ Interpretation: Equation(1) states that the probability of event E is a weighted average of the conditional probability of E given that F occurs and the conditional probability of E given that F does not occur.

▶ Each conditional probability is weighted by the probability of the event on which it is conditioned.

Rule of total probability

Suppose that events $F_1, F_2, \ldots, F_k$, are mutually exclusive and exhaustive; that is, exactly one of the events must occur. Then for any event E,

$$P(E) = P(E|F_1)P(F_1) + P(E|F_2)P(F_2) + \ldots + P(E|F_k)P(F_k)$$

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

Example 1: Application- Insurance policy

- An insurance company[1] believes that people can be divided
  into two classes—those who are prone to have accidents and
  those who are not. The data indicate that an accident-prone
  person will have an accident in a 1-year period with probability
  0.1; the probability for all others is 0.05. Suppose that the
  probability is 0.2 that a new policyholder is accident-prone.
  What is the probability that a new policyholder will have an
  accident in the first year?

---

[1]Ross, Sheldon M. Introductory statistics. Academic Press, 2017.

Application- Insurance policy-solution

▶ Define events

  ▶ E: A new policy holder will have an accident in the first year.
  ▶ F: A new policy holder is accident prone.

▶ Given

  ▶ an accident-prone person will have an accident in a 1-year period with probability 0.1, i.e., $P(E|F) = 0.1$
  ▶ the probability for all others is 0.05, i.e., $P(E|F^c) = 0.05$
  ▶ the probability is 0.2 that a new policyholder is accident-prone; $P(F) = 0.2$

▶ To compute probability that a new policyholder will have an accident in the first year, $P(E)$.

▶ $P(E) = P(F)P(E|F) + P(F^c)P(E|F^c) =$
  $0.2 \times 0.1 + 0.8 \times 0.05 = 0.06$

There is a 6 percent chance that a new policyholder will have an accident in the first year.

Bayes' rule

▶ Suppose we are now interested in the conditional probability of event F conditioned on E. We know

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

▶ From definition

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{P(E|F)P(F)}{P(F)P(E|F) + P(F^c)P(E|F^c)}$$

Bayes' rule: Suppose that events $F_1, F_2, \ldots, F_k$, are mutually exclusive and exhaustive; Then for any event E,

$$P(F_i|E) = \frac{P(E|F_i)P(F_i)}{\displaystyle\sum_{i=1}^{k} P(E|F_i)P(F_i)}$$

here E can be any one of events $F_1, F_2, \ldots, F_k$

Example 1: Application- Insurance policy

▶ An insurance company believes that people can be divided into two classes—those who are prone to have accidents and those who are not. The data indicate that an accident-prone person will have an accident in a 1-year period with probability 0.1; the probability for all others is 0.05. Suppose that the probability is 0.2 that a new policyholder is accident-prone. If a new policyholder has an accident in the first year, what is the probability that he or she is accident-prone?

Application- Insurance policy-solution

▶ Already computed probability that a new policyholder will
  have an accident in the first year, $P(E) =$
  $P(F)P(E|F) + P(F^c)P(E|F^c) = 0.2 \times 0.1 + 0.8 \times 0.05 = 0.06$

▶ Now "If a new policyholder has an accident in the first year",
  implies occurence of event E.

▶ What is the probability that he or she is accident-prone? In
  other words, what is
  P(accident prone|policy holder has an accident in first year).
  This is equivalent to $P(F|E)$

▶ Applying Bayes' rue we get
  $P(F|E) = \frac{P(F)P(E|F)}{P(F)P(E|F) + P(F^c)P(E|F^c)} = \frac{0.02}{0.06} = \frac{1}{3}$

Therefore, given that a new policyholder has an accident in the
first year, the conditional probability that the policyholder is prone
to accidents is $1/3$.

Section summary

1. Law of total probability
2. Bayes' rule

Random variable
    Example: Rolling a dice twice
    Example: Tossing a coin three times

Random Variable

- ▶ When a probability experiment is performed, often we are not interested in all the details of the experimental result, but rather are interested in the value of some numerical quantity determined by the result.
- ▶ For example, in rolling a dice twice, often we care about only their sum of outcomes and are not concerned about the values on the individual dice.
  - ▶ That is, we may be interested in knowing that the sum is 7 and may not be concerned over whether the actual outcome was $(1, 6), (2, 5), (3, 4), (4, 3), (5, 2),$ or $(6, 1)$.
- ▶ These quantities of interest, or, more formally, these real-valued functions defined on the sample space, are known as random variables.
- ▶ Because the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.

Rolling a dice: Sample space

▶ Experiment: Roll a dice twice

▶ The sample space for this experiment is

$$
S = \left\{
\begin{array}{l}
(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6),
\end{array}
\right\}
$$

▶ Consider the probabilities associated with the two questions

   1. Of the outcomes, how many outcomes will result in a sum of
      outcomes as 7?
   2. Of the outcomes, how many outcomes will have the smaller of
      the outcomes as 3?

▶ Notice, the experiment and sample space used to answer both
   the questions are the same.

- ▶ Let X denote the sum of outcomes of the two rolls.
- ▶ Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the outcome is taken as value of Y.

Outcome	X	Y	Outcome	X	Y	Outcome	X	Y
(1,1)	2	1	(3,1)	4	1	(5,1)	6	1
(1,2)	3	1	(3,2)	5	2	(5,2)	7	2
(1,3)	4	1	(3,3)	6	3	(5,3)	8	3
(1,4)	5	1	(3,4)	7	3	(5,4)	9	4
(1,5)	6	1	(3,5)	8	3	(5,5)	10	5
(1,6)	7	1	(3,6)	9	3	(5,6)	11	5
(2,1)	3	1	(4,1)	5	1	(6,1)	7	1
(2,2)	4	2	(4,2)	6	2	(6,2)	8	2
(2,3)	5	2	(4,3)	7	3	(6,3)	9	3
(2,4)	6	2	(4,4)	8	4	(6,4)	10	4
(2,5)	7	2	(4,5)	9	4	(6,5)	11	5
(2,6)	8	2	(4,6)	10	4	(6,6)	12	6

Sum of rolls of the dice

▶ Let X denote the sum of outcomes of the two rolls.

▶ X takes the values 2,3,4,5,6,7,8,9,10,11, and 12.

Value of X	Relevant event
2	$\{(1,1)\}$
3	$\{(1,2),(2,1)\}$
4	$\{(1,3),(2,2),(3,1)\}$
5	$\{(1,4),(2,3),(3,2),(4,1)\}$
⋮	⋮
9	$\{(3,6),(4,5),(5,4),(6,3)\}$
10	$\{(4,6),(5,5),(6,4)\}$
11	$\{(5,6),(6,5)\}$
12	$\{(6,6)\}$

▶ We say X is a random variable taking on one of the values
2,3,4,5,6,7,8,9,10,11, and 12 with respective probabilities

Probability of X	Probability of relevant event	Probability
$P\{X = 2\}$	$P(\{(1,1)\})$	$\frac{1}{36}$
$P\{X = 3\}$	$P(\{(1,2),(2,1)\})$	$\frac{2}{36}$
$P\{X = 4\}$	$P(\{(1,3),(2,2),(3,1)\})$	$\frac{3}{36}$
$P\{X = 5\}$	$P(\{(1,4),(2,3),(3,2),(4,1)\})$	$\frac{4}{36}$
$\vdots$	$\vdots$	$\vdots$
$P\{X = 9\}$	$P(\{(3,6),(4,5),(5,4),(6,3)\})$	$\frac{4}{36}$
$P\{X = 10\}$	$P(\{(4,6),(5,5),(6,4)\})$	$\frac{3}{36}$
$P\{X = 11\}$	$P(\{(5,6),(6,5)\})$	$\frac{2}{36}$
$P\{X = 12\}$	$P(\{(6,6)\})$	$\frac{1}{36}$

Lesser of the two values

▶ Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the out come is taken as value of Y.

▶ Y takes the values 1,2,3,4,5, and 6.

Y	Relevant event
1	$\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)\}$
2	$\{(2,2),(2,3),(2,4),(2,5),(2,6),(3,2),(4,2),(5,2),(6,2)\}$
3	$\{(3,3),(3,4),(3,5),(3,6),(4,3),(5,3),(6,3)\}$
4	$\{(4,4),(4,5),(4,6),(5,4),(6,4)\}$
5	$\{(5,5),(5,6),(6,5)\}$
6	$\{(6,6)\}$

► We say Y is a random variable taking on one of the values 1,2,3,4,5, and 6 with respective probabilities

Y	Relevant event	Probability
1	$\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)\}$	$\frac{11}{36}$
2	$\{(2,2),(2,3),(2,4),(2,5),(2,6),(3,2),(4,2),(5,2),(6,2)\}$	$\frac{9}{36}$
3	$\{(3,3),(3,4),(3,5),(3,6),(4,3),(5,3),(6,3)\}$	$\frac{7}{36}$
4	$\{(4,4),(4,5),(4,6),(5,4),(6,4)\}$	$\frac{5}{36}$
5	$\{(5,5),(5,6),(6,5)\}$	$\frac{3}{36}$
6	$\{(6,6)\}$	$\frac{1}{36}$

Tossing a coin three times: Sample space

- ▶ Experiment: Toss a coin three times.
- ▶ The sample space for this experiment is

  $$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

- ▶ Consider the probabilities associated with the two questions:
  1. Of the three tosses, how many tosses will be heads?
  2. Of the three tosses, which toss results in a heads first?, i.e first, second or third toss is a head?
- ▶ Notice the experiment and sample space used to answer both the questions are the same.

▶ Let X denote the number of heads that appear. Let Y denote the toss in which a head appears first.

Outcome	X	Y
HHH	3	1
HHT	2	1
HTH	2	1
HTT	1	1
THH	2	2
THT	1	2
TTH	1	3
TTT	0	NIL

Number of tosses that will be heads

▶ Let X denote the number of heads that appear.

▶ X takes the values 0,1,2,3

Value of X	Relevant event
0	$\{(TTT)\}$
1	$\{(HTT),(THT),(TTH)\}$
2	$\{(HHT),(HTH),(THH)\}$
3	$\{(HHH)\}$

▶ We say X is a random variable taking on one of the values 0, 1, 2, and 3 with respective probabilities

  ▶ $P\{X = 0\} = P\{(TTT)\} = \frac{1}{8}$
  ▶ $P\{X = 1\} = P\{(HTT),(THT),(TTH)\} = \frac{3}{8}$
  ▶ $P\{X = 2\} = P\{(HHT),(HTH),(THH)\} = \frac{3}{8}$
  ▶ $P\{X = 3\} = P\{(HHH)\} = \frac{1}{8}$

Which toss results in a heads first

▶ Let Y denote the toss in which a head appears first.

▶ Y takes the values 1,2,3, and NIL

Value of Y	Relevant event
1	$\{(HHH), (HHT), (HTH), (HTT)\}$
2	$\{(THH), (THT)\}$
3	$\{(TTH)\}$
NIL	$\{(TTT)\}$

▶ We say Y is a random variable taking on one of the values 1, 2, 3, and NIL with respective probabilities
  ▶ $P\{Y = 1\} = P\{(HHH), (HHT), (HTH), (HTT)\} = \frac{4}{8}$
  ▶ $P\{Y = 2\} = P\{(THH), (THT)\} = \frac{2}{8}$
  ▶ $P\{Y = 3\} = P\{(TTH)\} = \frac{1}{8}$
  ▶ $P\{Y = NIL\} = P\{(TTT)\} = \frac{1}{8}$

Section summary

1. Definition of a random variable and examples

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Example: Application- life insurance

▶ A life insurance agent has 2 elderly clients, each of whom has a life insurance policy that pays ₹1 lakh upon death.

▶ Let A be the event that the younger one dies in the following year, and let B be the event that the older one dies in the following year.

▶ Assume that A and B are independent, with respective probabilities $P(A) = .05$ and $P(B) = .10$.

▶ Let X denotes the total amount of money (in units of ₹lakhs) that will be paid out this year to any of these clients' beneficiaries.

▶ X is a random variable that takes on one of the possible values 0, 1, 2 with respective probabilities

Example: Application- life insurance

- ► Let X denote the total amount of money in ₹lakhs disbursed.
- ► X takes the values 0,1,2.

Value of X	Relevant event
0	$A^c \cap B^c$
1	$(A \cap B^c) \cup (A^c \cap B)$
2	$A \cap B$

- ► We say X is a random variable taking on one of the values 0, 1, and 2 with respective probabilities
  - ► $P\{X = 0\} = P(A^c \cap B^c) = 0.95 \times 0.9 = 0.855$
  - ► $P\{X = 1\} = P((A \cap B^c) \cup (A^c \cap B)) = (0.05 \times 0.9) + (0.95 \times 0.1) = 0.140$
  - ► $P\{X = 2\} = P(A \cap B) = 0.05 \times 0.1 = 0.005$

Discrete and Continuous random variables

Definition
A random variable that can take on at most a countable number of possible values is said to be a discrete random variable.

- ▶ Thus, any random variable that can take on only a finite number or countably infinite number of different values is discrete.

- ▶ There also exist random variables whose set of possible values is uncountable.

Definition
When outcomes for random event are numerical, but cannot be counted and are infinitely divisible, we have continuous random variables.

Section summary

- ▶ What is a random variable.
- ▶ Probability of a random variable.
- ▶ Defined discrete and continuous random variable.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Dicrete and continuous random variable

Discrete and Continuous random variables

Definition
A random variable that can take on at most a countable number of possible values is said to be a discrete random variable.

- ▶ Thus, any random variable that can take on only a finite number or countably infinite number of different values is discrete.
- ▶ There also exist random variables whose set of possible values is uncountable.

Definition
When outcomes for random event are numerical, but cannot be counted and are infinitely divisible, we have continuous random variables.

Discrete and continuous random variable

- A discrete random variable is one that has possible values that are discrete points along the real number line.
  - Discrete random variables typically involve counting.
- A continuous random variable is one that has possible values that form an interval along the real number line.
  - Continuous random variables typically involve measuring.

Example: Apartment complex

Apartment complex data:

▶ There are four floors in the apartment complex.

▶ Each floor has three apartments: a one bedroom, a two bedroom and a three bedroom apartment.

▶ The data on the apartments is summarized in the table

Apartment complex data

Apartment number	Floor number	No. of bedrooms	Size of apartment (sq.ft)	Distance of apartment from lift (meters)
1	1	1	900.23	503.5
2	1	2	1175.34	325.6
3	1	3	1785.85	450.8
4	2	1	900.48	500.1
5	2	2	1175.23	324.5
6	2	3	1785.35	456.7
7	3	1	900.53	502.5
8	3	2	1176.34	325.6
9	3	3	1787.85	450.8
10	4	1	900.78	500.1
11	4	2	1176.03	325.4
12	4	3	1784.85	455.7

Apartment complex

▶ Random experiment: Randomly selecting an apartment in an apartment complex of 12 apartments.

▶ $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

Questions

1. Let the random variable be number of bedrooms, what are the possible values that might be observed?
   Answer: 1,2,3

2. Let the random variable be floor number of the apartment. What are the possible values that might be observed?
   Answer: 1, 2,3,4

3. Let the random variable be size of the apartment. What are the possible values that might be observed?
   Answer: [900,1800] sq. ft

4. Let the random variable be distance of the apartment from the lift. What are the possible values that might be observed?
   Answer: [324,505] meters

Discrete versus continuous

- ▶ Which variables are discrete random variables?
    - ▶ Number of bedrooms, floor number.
- ▶ Which variables are continuous random variables?
    - ▶ Size, distance to the lift.

Discrete and continuous- more examples

▶ Discrete:
  ▶ Number of people in a household
  ▶ Number of languages a person can speak
  ▶ Number of times a person takes a particualr test before qualifying.
  ▶ Number of accidents in an intersection.
  ▶ Number of spelling mistakes in a report.
▶ Continuous:
  ▶ Temperature of a person.
  ▶ Height of a person.
  ▶ Speed of a vehicle.
  ▶ Time taken by a person to write an exam.

Section summary

- ▶ Definitions of Discrete random variable versus continuous random variable
- ▶ Identify discrete and continuous random variables.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Probability mass function, graph, and examples
Probability mass function

Probability mass function (p.m.f)

- ▶ A random variable that can take on at most a countable number of possible values is said to be a discrete random variable.

- ▶ Let X be a discrete random variable, and suppose that it has n possible values, which we will label $x_1, x_2, \ldots, x_n$.

- ▶ For a discrete random variable X, we define the probability mass function $p(x)$ of X by

$$p(x_i) = P(X = x_i)$$

- ▶ Represent it in tabular form

X	x_1	x_2	x_3	$\ldots$	$\ldots$	x_n
$P(X = x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	$\ldots$	$\ldots$	$p(x_n)$

Properties of p.m.f

▶ The probability mass function $p(x)$ is positive for at most a countable number of values of x. That is, if X must assume one of the values $x_1, x_2, \ldots$, then

  1. $p(x_i) \geq 0, i = 1, 2, \ldots$
  2. $p(x) = 0$ for all other values of x

▶ Represent it in tabular form

X	x_1	x_2	x_3			
$P(X = x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$			

▶ Since X must take one of the values x_i, we have

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

Example

- Suppose X is a random variable that takes three values, 0, 1, and 2 with probabilities
  - $p(0) = P(X = 0) = \frac{1}{4}$
  - $p(1) = P(X = 1) = \frac{1}{2}$
  - $p(2) = P(X = 2) = \frac{1}{4}$
- Tabular form

X	0	1	2
$P(X = x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

- Verify that $\displaystyle\sum_{i=1}^{3} p(x_i) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

Statistics for Data Science -1
└─Probability mass function, graph, and examples
  └─Probability mass function

Example

Let X be a random variable that takes values 1,2,3,4,5. Which of the following are probability mass functions?

1.

X	1	2	3	4	5
$P(X = x_i)$	0.4	0.1	0.2	0.1	0.3

NO

2.

X	1	2	3	4	5
$P(X = x_i)$	0.2	0.3	0.4	-0.1	0.2

NO

3.

X	1	2	3	4	5
$P(X = x_i)$	0.3	0.1	0.2	0.4	0.0

YES

Example

- Suppose X is a random variable that takes values, 0, 1, 2, . . . with probabilities
  - $p(i) = c\frac{\lambda^i}{i!}$, for some positive λ
- What is the value of c?

  - $\sum_{i=0}^{\infty} p(x_i) = 1$

  - $\sum_{i=0}^{\infty} c\frac{\lambda^i}{i!} = 1$

  - $c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$

  - Recall, $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$, hence $c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = ce^\lambda$

  - Hence, $c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = ce^\lambda = 1$ which gives $c = e^{-\lambda}$

Example: Rolling a dice twice

- $S = \left\{ \begin{array}{l} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6), \end{array} \right\}$

- X is a random variable which is defined as sum of outcomes

  - Probability mass function

X	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

  - Verify: $\displaystyle\sum_{i=1}^{11} p(x_i) = \frac{36}{36} = 1$

- Y is the random variable which takes the lesser of the values of the outcomes

  - Probability mass function

Y	1	2	3	4	5	6
$P(Y = y_i)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

  - Verify: $\displaystyle\sum_{i=1}^{6} p(y_i) = \frac{36}{36} = 1$

Statistics for Data Science -1
└─ Probability mass function, graph, and examples
  └─ Probability mass function

Example: Tossing a coin three times

- $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- X is the random variable which counts the number of heads in the tosses

  - Probability mass function

X	0	1	2	3
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

  - Verify: $\displaystyle\sum_{i=1}^{4} p(x_i) = \frac{8}{8} = 1$

- Y is the random variable which counts the toss in which heads appears first

  - Probability mass function

Y	1	2	3	NIL
$P(Y = y_i)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

  - Verify: $\displaystyle\sum_{i=1}^{4} p(y_i) = \frac{8}{8} = 1$

Section summary

- ▶ Probability mass function.
- ▶ Properties of probability mass function.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Probability mass function, graph, and examples
    Probability mass function
    Graph of probability mass function

Graph of probability mass function

▶ It is helpful to illustrate the probability mass function in a
  graphical format by plotting $P(X = x_i)$ on the y-axis against
  x_i on the x-axis.

▶ Let's look at a few examples

Example

X	0	1	2
$P(X = x_i)$	0.25	0.5	0.25
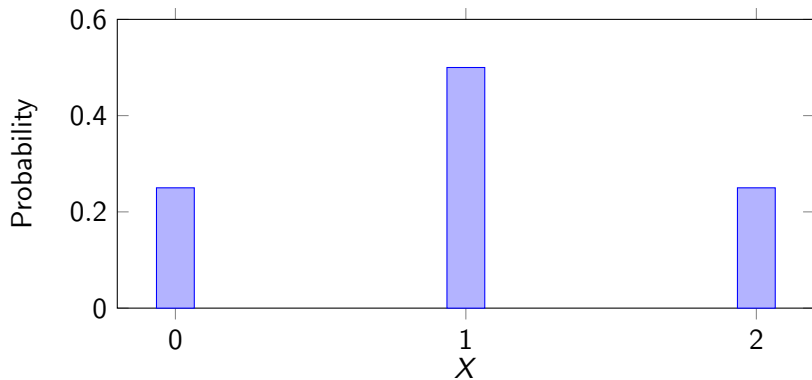
Example

X	0	1	2
$P(X = x_i)$	0.5	0.25	0.25

Statistics for Data Science -1
└─ Probability mass function, graph, and examples
  └─ Graph of probability mass function
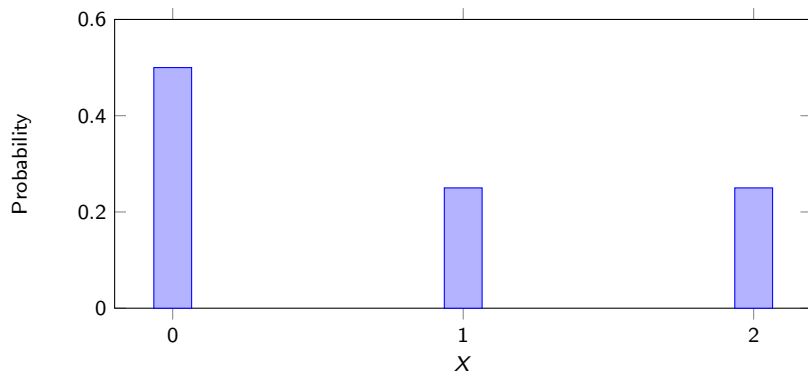
Example

X	0	1	2
$P(X = x_i)$	0.35	0.55	0.10
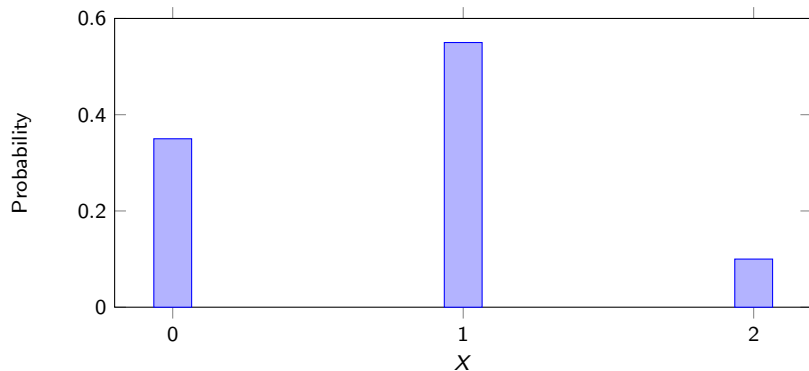
Example: positive skewed distribution

X	1	2	3	4
$P(X = x_i)$	0.4	0.3	0.2	0.1

Statistics for Data Science -1
└─ Probability mass function, graph, and examples
   └─ Graph of probability mass function
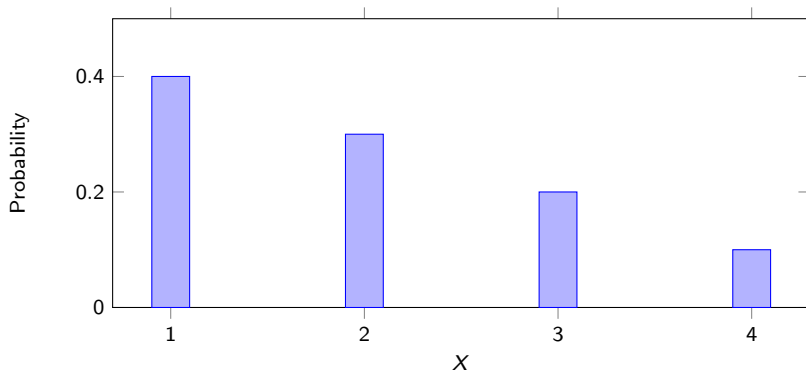
Example: negative skewed distribution

X	1	2	3	4
$P(X = x_i)$	0.1	0.2	0.3	0.4
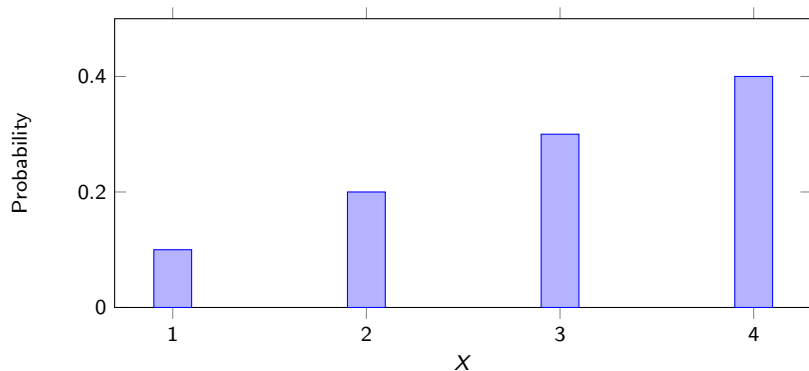
Rolling a dice once: X outcome

X	1	2	3	4	5	6
$P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
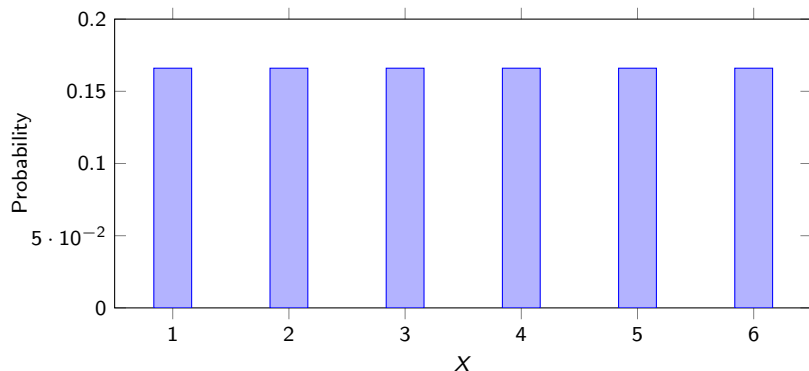
Rolling a dice twice: X-sum of outcomes

X	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Statistics for Data Science -1
└ Probability mass function, graph, and examples
  └ Graph of probability mass function

Rolling a dice twice: X-smaller of outcomes

X	1	2	3	4	5	6
$P(X = x_i)$	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$
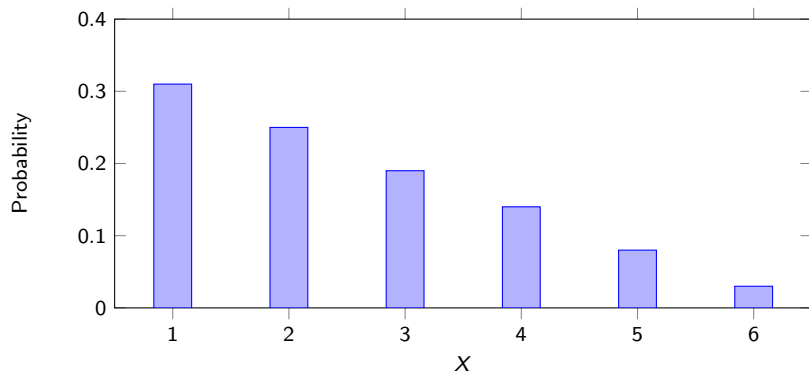
Toss a coin once: X- outcome

X	0	1
$P(X = x_i)$	0.5	0.5
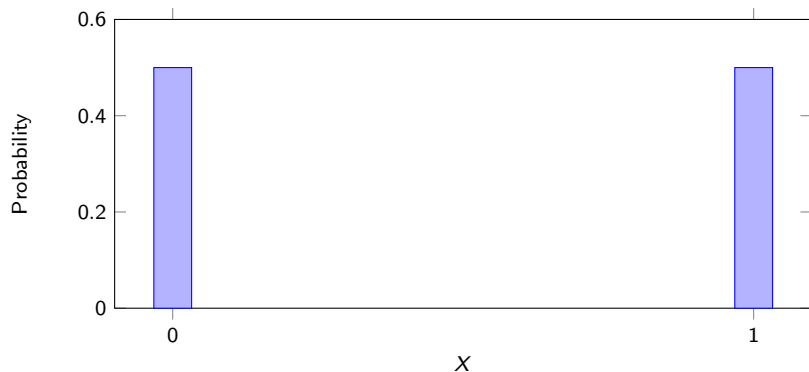
Tossing a coin thrice: X-number of heads

X	0	1	2	3
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
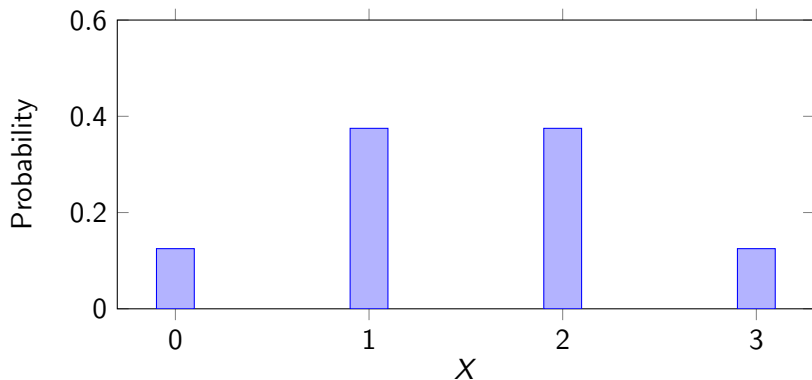
Tossing a coin thrice: X-toss head appearing first

X	1	2	3	NIL
$P(X = x_i)$	$\frac{4}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
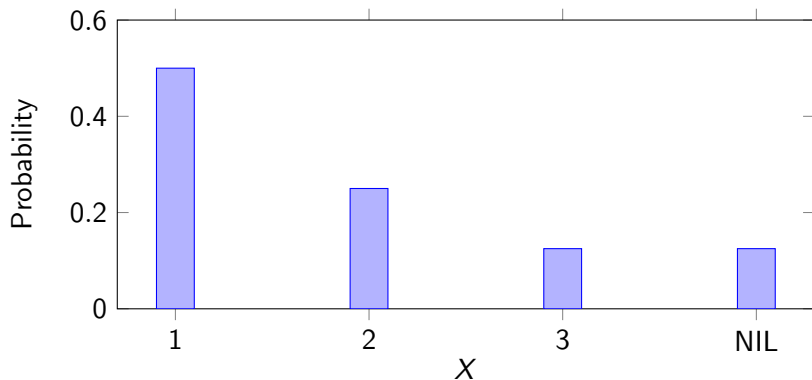
Section summary

► Graph of p.m.f and examples

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Probability mass function, graph, and examples
    Probability mass function
    Graph of probability mass function

Cumulative distribution function, graph, and examples

Cumulative distribution function

▶ The cumulative distribution function (cdf), F, can be expressed by

$$F(a) = P(X \leq a)$$

▶ If X is a discrete random variable whose possible values are $x_1, x_2, x_3, \ldots$, where $x_1 < x_2 < x_3 \ldots$, then the distribution function F of X is a step function.

Step function

▶ Let X be a discrete random variable with the following probability mass function.

X	1	2	3	4
$P(X = x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$

▶ The cumulative distribution function of X is given by

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a \end{cases}$$

Note that the size of the step at any of the values 1, 2, 3, and 4 is equal to the probability that X assumes that particular value.
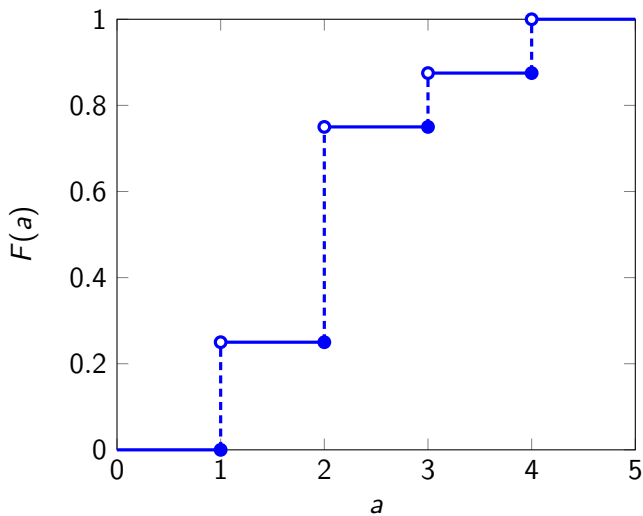
Section summary

- ▶ Probability mass function- tabular form and graph.
- ▶ Cumulative distribution function- definition and graph.
- ▶ Key ideas:
    - ▶ Shape of distribution: skewed, symmetric, constant, etc.
    - ▶ Answer questions about distribution of random variable.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Application: Credit cards

Introduction

We are interested in analyzing the number of credit cards owned by people. Towards this, we collect data on a number of people and ask them how many credit cards they own. Their response is recorded.

► Random experiment: Consider the random experiment of selecting an adult at random from the sample.

► Random variable: The number of credit cards owned by the person.

  ► This is a discrete random variable.

► The following table summarizes the probability distribution of the number of credit cards per person based on the relative frequencies.

►

X	0	1	2	3	4
$P(X = x_i)$	0.42	0.36	0.14	0.06	0.02

Probability mass function
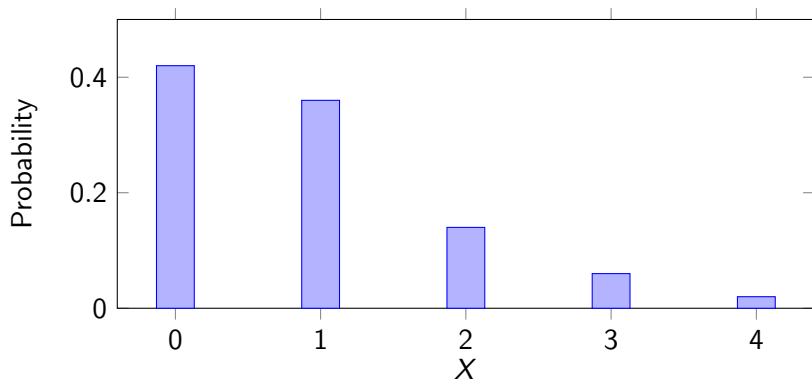
Questions

- Describe the distribution
  - The distribution is skewed right with a peak at 0.
  - Number of credit cards owned by people vary between 0 to 4 credit cards.
- Choose an adult at random. Is he/she more likely to have 0 credit cards or 2 or more credit cards?
  - The probability that an adult has no credit cards is 0.42, while the probability of having 2 or more credit cards is about 0.22, so the probability of having no credit cards is higher.

Questions-continued

- You take a random sample of 500 people and ask them how many credit cards they own. Would you be surprised at the following:
  - Everyone owns a credit card.
    YES: 42% of adults do not own credit cards. Hence, it is unlikely that every one of the 500 would own credit cards.
  - 72 people respond that they own two credit cards.
    NO: 14% own two credit cards. 14% of 500 = 70, so it is likely that 72 people from sample of 500 own two credit cards.
- Again choose an adult at random. How many credit cards would you "Expect" that person to own?

Section summary

- ▶ Probability mass function- tabular form and graph.
- ▶ Cumulative distribution function- definition and graph.
- ▶ Key ideas:
  - ▶ Shape of distribution: skewed, symmetric, constant, etc.
  - ▶ Answer questions about distribution of random variable.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Expectation of a random variable

Introduction

Consider the following game of rolling a dice once.

▶ If the outcome is even- you lose an amount equal to the outcome

▶ If the outcome is odd- you win an amount equal to the outcome.

▶ In other words, the gains/losses are as per table

Outcome	1	2	3	4	5	6
Winning	+1	-2	+3	-4	+5	-6

*: A winning of $-x$ indicates a loss of x amount.

▶ Question: Would you play this game?

Simulating the game

▶ First, roll the dice 100 times. Observe the outcomes. They
  are summarised in the table.

1	5	4	6	4
6	4	3	1	5
5	2	1	6	2
6	5	4	5	5
2	4	2	2	5
4	1	3	4	5
6	1	5	2	5
1	6	6	5	2
3	6	5	6	3
5	3	2	5	4
3	3	5	4	4
5	1	2	3	4
3	2	1	1	6
1	3	4	3	4
4	4	5	6	1
3	3	5	1	1
1	6	5	4	3
5	1	4	6	5
4	4	4	3	6
5	5	4	1	3

▶ Rolling 100 times

Outcome	Winning	Frequency	Relative frequency
1	+1	16	0.16
2	-2	10	0.10
3	+3	16	0.16
4	-4	21	0.21
5	+5	23	0.23
6	-6	14	0.14
		100	1

Average winnings: -0.09

▶ Rolling a 1000 times

Outcome	Winning	Frequency	Relative frequency
1	+1	177	0.177
2	-2	177	0.177
3	+3	167	0.167
4	-4	153	0.153
5	+5	163	0.163
6	-6	163	0.163
		1000	1

Average winnings: -0.451

Observations

▶ The relative frequency of each of the six possible outcomes is close to the probability of $\frac{1}{6}$ for the respective outcomes.

▶ Hence, it suggests, that if I repeat rolling the dice for a very large number of times, our average gain should be

$$1\frac{1}{6} - 2\frac{1}{6} + 3\frac{1}{6} - 4\frac{1}{6} + 5\frac{1}{6} - 6\frac{1}{6} = -0.5$$

▶ This is close to what we got as the average winning for 1000 rolls of the dice.

Expectation of a random variable

Definition

Let X be a discrete random variable taking values $x_1, x_2, \ldots,$. The expected value of X denoted by $E(X)$ and referred to as Expectation of X is given by

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i)$$

▶ The Expectation of a random variable can be considered the "long-run-average" value of the random variable in repeated independent observations.

▶ Lets apply the definition to the examples we have considered before

Rolling a dice once

▶ Random experiment: Roll a dice once.

▶ Sample space: $S = \{1, 2, 3, 4, 5, 6\}$

▶ Random variable X is the outcome of the roll.

▶ The probability distribution is given by

X	1	2	3	4	5	6
$P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

▶ $E(X) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$.

▶ Does this mean that if we roll a dice once, should we expect the outcome to be 3.5?

▶ NO!!-the expected value tells us is what we would expect the average of a large number of rolls to be in the long run.

Summary of the rolling dice simulation

	100 rolls		1000 rolls		
Outcome	Freq	Rel. Freq	Freq	Rel. Freq	Probability
1	16	0.16	177	0.177	0.166667
2	10	0.1	177	0.177	0.166667
3	16	0.16	167	0.167	0.166667
4	21	0.21	153	0.153	0.166667
5	23	0.23	163	0.163	0.166667
6	14	0.14	163	0.163	0.166667
		3.67		**3.437**	**3.5**

▶ Notice that average of the rolls need not be exactly 3.5.

▶ However, we can expect it to be close to 3.5.

▶ The expected value of X is a theoretical average.

Rolling a dice twice

▶ X is a random variable which is defined as sum of outcomes

▶ Probability mass function

X	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

▶ If I rolled two dice a large number of times, what can I expect the average of the sum of the outcomes to be?

$$E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + \ldots + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = 7$$

▶ Interpretation: When two dice are rolled over and over for a long time, the mean sum of the two dice is 7.
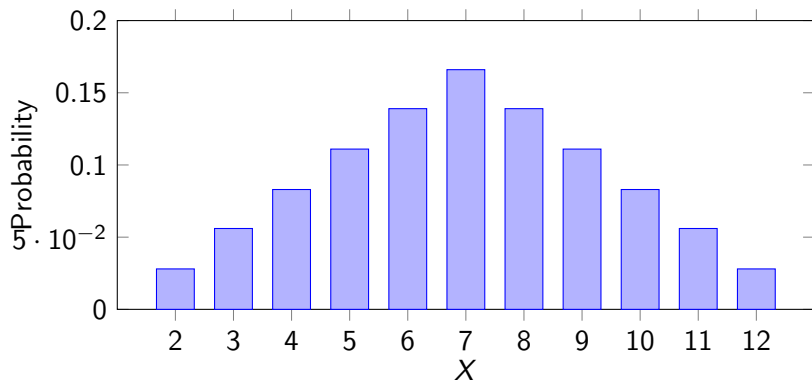
Tossing a coin thrice

- $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- X is the random variable which counts the number of heads in the tosses
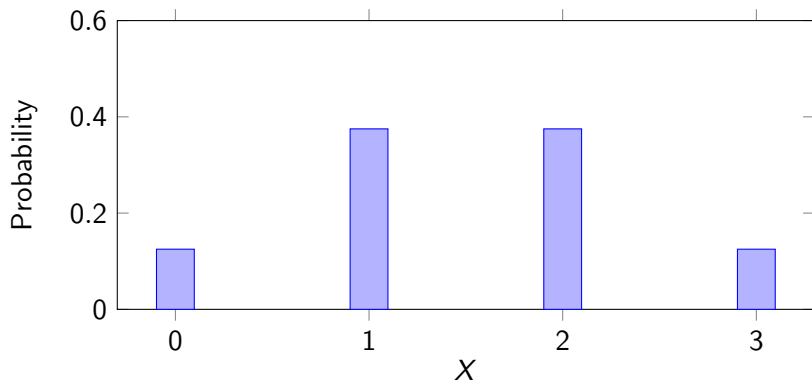
  - Probability mass function

X	0	1	2	3
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

  - $E(X) = \sum_{i=0}^{3} x_i p(x_i) =$

    $\dfrac{(0 \times 1) + (1 \times 3) + (2 \times 3) + (3 \times 1)}{8} = \dfrac{3}{2}$

- Interpretation: When a coin is tossed three times over and over for a long time, the mean number of heads in the three tosses is 1.5.

Bernoulli random variable

▶ A random variable that takes on either the value 1 or 0 is called a Bernoulli random variable.

▶ Let X be a Bernoulli random variable that takes on the value 1 with probability p.

▶ The probability distribution of the random variable is

X	0	1
$P(X = x_i)$	$1 - p$	p

▶ Expected value of a Bernoulli random varaible:

$$E(X) = 0 \times (1 - p) + 1 \times p = p$$

Discrete uniform random variable

▶ Let X be a random variable that is equally likely to takes any of the values $1, 2, \ldots, n$

▶ Probability mass function

X	1	2	$\ldots$	n
$P(X = x_i)$	$\frac{1}{n}$	$\frac{1}{n}$	$\ldots$	$\frac{1}{n}$

▶ $E(X) = \sum_{i=1}^{n} x_i p(x_i) = \frac{(1 \times 1) + (2 \times 1) + \ldots + (n \times 1)}{n} =$
$\frac{n(n+1)}{2 \times n} = \frac{(n+1)}{2}$

Section summary

- ▶ Notion of expectation
- ▶ Bernoulli and Discrete uniform random variable.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Properties of expectation

Expectation of a function of a random variable

Proposition

Let X be a discrete random variable which takes values x_i along with its probability mass function, $P(X = x_i)$. Let g be any real values function, The expected value of $g(X)$ is

$$E(g(X)) = \sum_i g(x_i)P(X = x_i)$$

Corollary

If a and b are constants, $E(aX + b) = aE(X) + b$

Example

- Let X be a discrete random variable with the following distribution

X	-1	0	1
$P(X = x_i)$	0.2	0.5	0.3

- Let $Y = g(X) = X^2$. What is $E(Y)$?
- $E(Y) = (-1^2) \times 0.2 + 0 \times 0.5 + 1^2 \times 0.3 = 0.5$
- Distribution of Y

Y	0	1
$P(Y = y_i)$	0.5	0.5

- NOTE: $0.5 = E(X^2) \neq (E(X))^2 = 0.01$

Example

Sanjay and Anitha work for the same company. Anitha's Diwali bonus is a random variable whose expected value is ₹15,000.

▶ Sanjay's bonus is set to equal 90 percent of Anita's, find the expected value of Sanjay's bonus.
  ▶ Let X denote Anita's bonus. Given $E(X) = 15,000$.
  ▶ Let Sanjay's bonus be Y. Given $Y = 0.9X$
  ▶ Hence $E(Y) = 0.9 \times E(X) = $₹13,500

▶ If Sanjay's is set to equal ₹1000 more than Anitha's, find his expected bonus.
  ▶ In this case, $Y = X + 1000$
  ▶ Hence $E(Y) = E(X) + 1000 = $ ₹16,000

Expectation of sum of two random variables

▶ The expected value of the sum of random variables is equal to
the sum of the individual expected values. i.e Let X and Y be
two random variables. Then,

$$E(X + Y) = E(X) + E(Y)$$

Example: Rolling a dice

▶ Let X be the outcome of a fair dice. Let Y be the outcome of another fair dice.

▶ We know $E(X) = E(Y) = 3.5$

▶ $X + Y$ is the sum of outcomes of both the dice rolled together. Then,

$$E(X + Y) = E(X) + E(Y) = 3.5 + 3.5 = 7$$

This is the same expectation of the sum of outcomes of rolling a dice twice.

Hypergeometric random variable

- Suppose that a sample of size n is to be chosen randomly (without replacement) from a box containing N balls, of which m are red and $N - m$ are blue.

- Let X denote the number of red balls selected, then

$$P(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, i = 0, 1, 2, \ldots, n$$

- X is a said to be a hypergeometric variable for some values of n, m, and N

- $E(X) = \frac{nm}{N}$

Example

Two students are randomly chosen from a group of 20 boys and 10 girls. Let X denote the number of boys chosen, and let Y denote the number of girls chosen.

1. Find $E(X)$.
   - X is a Hypergeometric rv. $N = 30, m = 20, n = 2$. Hence $E(X) = \frac{2 \times 20}{30} = \frac{4}{3}$
2. Find $E(Y)$.
   - Y is a Hypergeometric rv. $N = 30, m = 10, n = 2$. Hence $E(Y) = \frac{2 \times 10}{30} = \frac{2}{3}$
3. Find $E(X + Y)$.
   - $E(X + Y) = E(X) + E(Y) = \frac{4}{3} + \frac{2}{3} = 2$

Expectation of sum of many random variables

▶ The result that the expected value of the sum of random variables is equal to the sum of the expected values holds for not only two but any number of random variables.

▶ Let $X_1, X_2, \ldots, X_k$ be k discrete random variables. Then,

$$E\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} E(X_i)$$

Example: Tossing a coin three times

- ▶ Toss a coin i times.
- ▶ Let X_i be a random variable which equals 1 if the outcome is a head, 0 otherwise.
- ▶ $E(X_i) = 0.5$
- ▶ $X_1 + X_2 + \ldots + X_n$ is the total number of heads in n tosses of the coin.
- ▶ $E(X_1 + X_2 + \ldots + X_n) = \displaystyle\sum_{i=1}^{n} E(X_i) = 0.5 \times n$
- ▶ For $n = 3, X_1 + X_2 + X_3$ is equal to the number of heads in three tosses of a coin.

$$E(X_1 + X_2 + X_3) = 3 \times 0.5 = 1.5$$

This is the same expectation of number of heads in three tosses of a coin.

Section summary

- ▶ Notion of expected value as long-run average.
- ▶ How to compute expected value of a random variable.
- ▶ Expectation of a function of a random variable
- ▶ Expectation of sums of random variables. .

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Introduction

▶ The expected value of a random variable gives the weighted average of the possible values of the random variable, it does not tell us anything about the variation, or spread, of these values.

▶ For instance, consider random variables X, Y, and Z, whose values and probabilities are as follows:

▶ $X = 0$ with probability 1

▶ $Y = \begin{cases} -2 & \text{with probability } \frac{1}{2} \\ 2 & \text{with probability } \frac{1}{2} \end{cases}$

▶ $Z = \begin{cases} -20 & \text{with probability } \frac{1}{2} \\ 20 & \text{with probability } \frac{1}{2} \end{cases}$

▶ $E(X) = E(Y) = E(Z) = 0$. However, we notice spread of Z is greater than spread of Y which is greater than spread of X

▶ Need for a measure of spread.

Variance of a random variable

▶ Let's denote expected value of a random variable X by the greek alphabet μ.

Definition

Let X be a random variable with expected value μ, then the variance of X, denoted by $Var(X)$ or $V(X)$, is defined by

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

▶ In other words, the Variance of a random variable X measures the square of the difference of the random variable from its mean, μ, on the average.

Computational formula for $Var(X)$

- $Var(X) = E(X - \mu)^2$
- $(X - \mu)^2 = X^2 - 2X\mu + \mu^2$
- Using properties of expectation we know
  $E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2$ which is same as
  $E(X^2) - \mu^2$
- Let's compute the variance of the random variables discussed earlier.

Rolling a dice once

- ▶ Random experiment: Roll a dice once.
- ▶ Sample space: $S = \{1, 2, 3, 4, 5, 6\}$
- ▶ Random variable X is the outcome of the roll.
- ▶ The probability distribution is given by

X	1	2	3	4	5	6
X^2	1	4	9	16	25	36
$P(X = x_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- ▶ $E(X) = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$.
- ▶ $E(X^2) = 1\frac{1}{6} + 4\frac{1}{6} + 9\frac{1}{6} + 16\frac{1}{6} + 25\frac{1}{6} + 36\frac{1}{6} = 15.167$
- ▶ $Var(X) = 15.167 - 3.5^2 = 2.917$

Rolling a dice twice

- ▶ X is a random variable which is defined as sum of outcomes
- ▶ Probability mass function

X	2	3	4	5	6	7	8	9	10	11	12
X^2	4	9	16	25	36	49	64	81	100	121	144
$P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

- ▶ $E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + \ldots + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = 7$
- ▶ $E(X^2) = 4 \times \frac{1}{36} + 9 \times \frac{2}{36} + \ldots + 121 \times \frac{2}{36} + 144 \times \frac{1}{36} = 54.833$
- ▶ $Var(X) = 54.833 - 49 = 5.833$
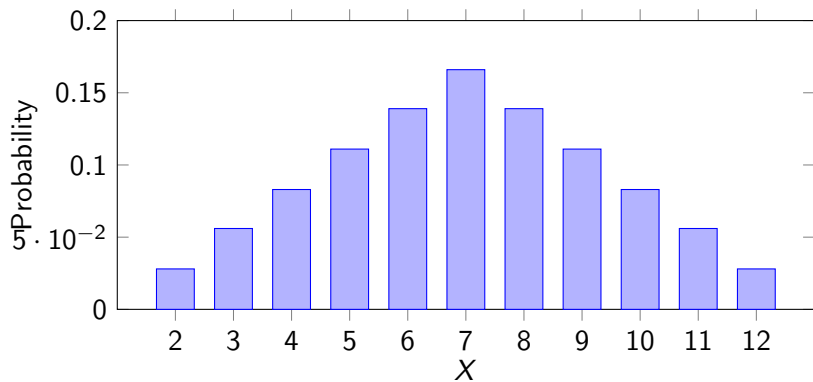
Tossing a coin thrice

- $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- X is the random variable which counts the number of heads in the tosses

- Probability mass function

X	0	1	2	3
X^2	0	1	4	9
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

- $E(X) = \sum_{i=0}^{3} x_i p(x_i) =$
  $$\frac{(0 \times 1) + (1 \times 3) + (2 \times 3) + (3 \times 1)}{8} = \frac{3}{2}$$

- $E(X^2) = \sum_{i=0}^{3} x_i p(x_i) =$
  $$\frac{(0 \times 1) + (1 \times 3) + (4 \times 3) + (9 \times 1)}{8} = \frac{24}{8} = 3$$

- $Var(X) = 3 - 2.25 = 0.75$

Bernoulli random variable

▶ A random variable that takes on either the value 1 or 0 is called a Bernoulli random variable.

▶ Let X be a Bernoulli random variable that takes on the value 1 with probability p.

▶ The probability distribution of the random variable is

X	0	1
X^2	0	1
$P(X = x_i)$	$1 - p$	p

▶ Expected value of a Bernoulli random varaible:
$E(X) = 0 \times (1 - p) + 1 \times p = p$

▶ Variance of a Bernoulli random variable:
$Var(X) = p - p^2 = p(1 - p)$

Discrete uniform random variable

- ▶ Let X be a random variable that is equally likely to takes any of the values $1, 2, \ldots, n$
- ▶ Probability mass function

X	1	2	...	n
X^2	1	4	...	n^2
$P(X = x_i)$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

- ▶ $E(X) = \frac{(n+1)}{2}$
- ▶ $E(X^2) = \frac{(n+1)(2n+1)}{6}$
- ▶ $Var(X) = \frac{n^2-1}{12}$

Section summary

- Definition of variance
- Computational formula of variance of a random variable.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Variance of a function of a random variable

Proposition

Let X be a random variable, let c be a constant, then

- $Var(cX) = c^2 Var(X)$
- $Var(X + c) = Var(X)$

Corollary

If a and b are constants, $V(aX + b) = a^2 Var(X)$

Proof.

We know $E(aX + b) = a\mu + b$. Hence,
$Var(aX + b) = E(aX + b - a\mu - b)^2 = E\left(a^2(X - \mu)^2\right) = a^2 E(X - \mu)^2 = a^2 Var(X)$ □

Variance of sum of two random variables

▶ The expected value of the sum of random variables is equal to the sum of the individual expected values. i.e if X and Y be two random variables. Then, $E(X + Y) = E(X) + E(Y)$.

▶ What can be said about the Variance of sum of two random variables?

▶ $Var(X + X) = Var(2X) = 4Var(X) \neq Var(X) + Var(X)$

Independent random variables

Definition

Random variables X and Y are independent if knowing the value of one of them does not change the probabilities of the other.

Example:

- ▶ Roll a dice twice. $S = \{(1,1), \ldots, (6,6)\}$
- ▶ X is the outcome of the first dice. Y be the out come of the second dice.
- ▶ Knowing $X = i$ does not change the probability of Y taking any value $1, 2, \ldots, 6$.
- ▶ X and Y are independent random variables.

Variance of sum of independent random variables

Result
Let X and Y be independent random variables. Then

$$Var(X + Y) = Var(X) + Var(Y)$$

Example: Rolling a dice twice

▶ Let X be the outcome of a fair dice. Let Y be the outcome of another fair dice.

▶ We know $E(X) = E(Y) = 3.5$

▶ $X + Y$ is the sum of outcomes of both the dice rolled together. Then, we know $E(X + Y) = E(X) + E(Y) = 3.5 + 3.5 = 7$.

▶ We also know $Var(X) = Var(Y) = 2.917$

▶ X and Y are independent, hence, $Var(X + Y) = 2.917 + 2.917 \approx 5.83$, which is the same as what we obtained earlier applying the computational formula.

Hypergeometric random variable

▶ Suppose that a sample of size n is to be chosen randomly (without replacement) from a box containing N balls, of which m are red and $N - m$ are blue.

▶ Let X denote the number of red balls selected, then

$$P(X = i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, i = 0, 1, 2, \ldots, n$$

▶ X is a said to be a hypergeometric variable for some values of n, m, and N

▶ $E(X) = \frac{nm}{N}$

▶ It can be verified that $Var(X) = \frac{nm}{N}\left[\frac{(n-1)(m-1)}{(N-1)} + 1 - \frac{nm}{N}\right]$

Variance of sum of many independent random variables

▶ The result that the variance of the sum of independent random variables is equal to the sum of the variances holds for not only two but any number of random variables.

▶ Let $X_1, X_2, \ldots, X_k$ be k discrete random variables. Then,

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

Example: Tossing a coin three times

▶ Toss a fair coin i times.

▶ Let X_i be a random variable which equals 1 if the outcome is a head, 0 otherwise.

▶ $E(X_i) = 0.5$, $Var(X_i) = 0.25$

▶ $X_1 + X_2 + \ldots + X_n$ is the total number of heads in n tosses of the coin.

▶ $Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} Var(X_i) = 0.5 \times n$

▶ For $n = 3$, $X_1 + X_2 + X_3$ is equal to the number of heads in three tosses of a coin.

$$Var(X_1 + X_2 + X_3) = 3 \times 0.25 = 0.75$$

This is the same as variance of number of heads in three tosses of a coin.

Section summary

▶ Properties of variance.

▶ Variance of sum of independent random variables.

Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.

Standard deviation of a random variable

Definition
The quantity $SD(X) = \sqrt{Var(X)}$ is called the standard deviation of X.

Hence, the standard deviation (SD) is the positive square root of the variance.

Remark
The standard deviation, like the expected value, is measured in the same units as is the random variable.

Properties of standard deviation

Proposition

Let X be a random variable, let c be a constant, then

▶ $SD(cX) = cSD(X)$

▶ $SD(X + c) = SD(X)$

1. If $Var(X) = 4$, what is $SD(3X)$? Answer: 6.
2. If $Var(2X + 3) = 16$, what is $SD(X)$? Answer: 2.

Application: family bonus

Sanjay and Anitha are a married couple who work for the same company. Anitha's Diwali bonus is a random variable whose expected value is ₹15,000 and standard deviation is ₹3,000. Sanjay's bonus is is a random variable whose expected value is ₹20,000 and standard deviation is ₹4,000. Assume the earnings of Sanjay and Anitha are independent of each other. What is the expected value and standard deviation of the total family bonus.

- ▶ Let X denote Anita's bonus. Given $E(X) = 15,000, SD(X) = 3,000$.
- ▶ Let Sanjay's bonus be Y. Given $E(Y) = 20,000, SD(Y) = 4,000$
- ▶ $E(X + Y) = E(X) + E(Y) =$ ₹35,000
- ▶ $SD(X + Y) = \sqrt{Var(X) + Var(Y)} =$ ₹5,000

Application: Lawyer's fees

- ▶ A lawyer must decide whether to charge a fixed fee of ₹25,000 or to take a contingency fee of ₹50,000 if she wins the case (and ₹0 if she loses).
- ▶ She estimates that her probability of winning is 0.5 (equal chance of winning or losing).
- ▶ Determine the expectation and standard deviation of her fee if
  - a She charges a fixed fee. $E(X) = 25,000, SD(X) = 0$
  - b She charges a contingency fee.
    $E(X) = 25,000, SD(X) = 25,000$

Section summary

- ▶ Notion of standard deviation of a random variable.
- ▶ Properties of standard deviation.
- ▶ Applications.

Learning objectives

1. Derive the formula for the probability mass function for Binomial distribution.
2. Understand the effect of parameters n and p on the shape of the Binomial distribution.
3. Expectation and variance of the binomial distribution.
4. To understand situations that can be modeled as a Binomial distribution.

Bernoulli trial

- ▶ A trial, or an experiment, whose outcome can be classified as either a success or a failure is called a Bernoulli trial.
- ▶ The sample space $S = \{Success, Failure\}$
- ▶ Let X be a random variable that takes the value 1 if the outcome is a success and value 0 if the outcome is 0.
- ▶ X is called a Bernoulli random variable.

Examples of Bernoulli trials

▶ Experiment: Tossing a coin: $S = \{Head, Tail\}$
  ▶ Success: Head
  ▶ Failure: Tail
▶ Experiment: Rolling a dice: $S = \{1, 2, 3, 4, 5, 6\}$
  ▶ Success: Getting a six.
  ▶ Failure: Getting any other number.
▶ Experiment: Opinion polls: $S = \{Yes, No\}$
  ▶ Success: Yes
  ▶ Failure: No
▶ Experiment: Salesperson selling an object:
  $S = \{Sale, No\ sale\}$
  ▶ Success: Sale
  ▶ Failure: No sale
▶ Experiment: Testing effectiveness of a drug:
  $S = \{Effective, Not\ effective\}$
  ▶ Success: Effective
  ▶ Failure: Not effective

Non Bernoulli trial

▶ Experiment: Randomly choosing a person and asking their age.

▶ Not Bernoulli- Outcomes are not 2.

Bernoulli random variable

▶ A random variable that takes on either the value 1 or 0 is called a Bernoulli random variable.

▶ X is a Bernoulli random variable that takes on the value 1 with probability p.

▶ The probability distribution of the random variable is

X	0	1
$P(X = x_i)$	$1 - p$	p

▶ Expected value of a Bernoulli random varaible:

$$E(X) = 0 \times (1 - p) + 1 \times p = p$$

▶ Variance of a Bernoulli random varaible:

$$V(X) = p - p^2 = p(1 - p)$$

Variance of Bernoulli Distribution

▶ The largest variance occurs when $p = \frac{1}{2}$, when success and failure are equally likely.

▶ In other words, the most uncertain Bernoulli trials, those with the largest variance, resemble tosses of a fair coin.

Section summary

- ► Bernoulli trial
- ► Bernoulli random variable

Learning objectives

1. Derive the formula for the probability mass function for Binomial distribution.
2. Understand the effect of parameters n and p on the shape of the Binomial distribution.
3. Expectation and variance of the binomial distribution.
4. To understand situations that can be modeled as a Binomial distribution.

Independent and identically distributed Bernoulli trials

▶ A collection of Bernoulli trials defines iid Bernoulli random variables, one for each trial.

▶ The abbreviation iid stands for independent and identically distributed.

Definition
A collection of random variables is iid if the random variables are independent and share a common probability distribution

Non Independent trials- Example

- ▶ Consider the experiment where three balls are chosen without replacement from a bag containing 20 red balls and 40 black balls. The number of red balls drawn is recorded. Is this a binomial experiment?

- ▶ NO!!- The balls are chosen without replacement. The colour of first ball will affect the chances of colour of the next balls- Independence criteria NOT satisfied.

Binomial random variable

- ▶ Suppose that n independent trials are performed, each of which results in either a "success" with probability p or a "failure" with probability $1 - p$.

- ▶ Let X is the total number of successes that occur in n trials, then X is said to be a binomial random variable with parameters n and p.

Non Binomial experiment

- ▶ Consider the experiment of rolling a dice until a 6 appears- is it a Binomial experiment?
- ▶ NO!!-Number of trials n is not fixed in this case

$n=3$ independent trials

► Let $n = 3$ independent Bernoulli trials.

► The outcomes of the independent trials are

S.No	Outcome
1	(s,s,s)
2	(s,s,f)
3	(s,f,s)
4	(s,f,f)
5	(f,s,s)
6	(f,s,f)
7	(f,f,s)
8	(f,f,f)

$n=3$ independent trials

▶ Let $n = 3$ independent Bernoulli trials.

▶ Let p is probability of success.

▶ The probabilities of outcomes of the independent trials are

S.No	Outcome	Probabilities
1	(s,s,s)	$p \times p \times p$
2	(s,s,f)	$p \times p \times (1 - p)$
3	(s,f,s)	$p \times (1 - p) \times p$
4	(s,f,f)	$p \times (1 - p) \times (1 - p)$
5	(f,s,s)	$(1 - p) \times p \times p$
6	(f,s,f)	$(1 - p) \times p \times (1 - p)$
7	(f,f,s)	$(1 - p) \times (1 - p) \times p$
8	(f,f,f)	$(1 - p) \times (1 - p) \times (1 - p)$

$n=3$ independent trials, $X =$ number of successes

- ▶ Let $n = 3$ independent Bernoulli trials.
- ▶ Let p is probability of success.
- ▶ $X=$ number of successes in 3 independent trials.
- ▶ The probabilities of outcomes of the independent trials are

S.No	Outcome	Number of successes	Probabilities
1	(s,s,s)	3	$p \times p \times p$
2	(s,s,f)	2	$p \times p \times (1 - p)$
3	(s,f,s)	2	$p \times (1 - p) \times p$
4	(s,f,f)	1	$p \times (1 - p) \times (1 - p)$
5	(f,s,s)	2	$(1 - p) \times p \times p$
6	(f,s,f)	1	$(1 - p) \times p \times (1 - p)$
7	(f,f,s)	1	$(1 - p) \times (1 - p) \times p$
8	(f,f,f)	0	$(1 - p) \times (1 - p) \times (1 - p)$

$n=3$ independent trials, $X =$ number of successes

- Let $n = 3$ independent Bernoulli trials.
- Let p is probability of success.
- Let $X =$ number of successes in 3 independent trials.
- The probability distribution of X

X	0	1	2	3
$P(X = i)$	$(1 - p)^3$	$3 \times p \times (1 - p)^2$	$3 \times p^2 \times (1 - p)$	p^3

n independent trials, $X =$ number of successes

- ▶ Let there be n independent Bernoulli trials.
- ▶ Let p is probability of success.
- ▶ $X =$ number of successes in n independent trials.
- ▶ The probabilities of outcomes of the independent trials are

S.No	Outcome	Number of successes	Probabilities
1	(s,s,...,s)	n	$p \times p \times \ldots \times p$
2	(s,s,...,f)	$n-1$	$p \times p \times \ldots \times (1-p)$
3	(s,...,f,s)	$n-1$	$p \ldots \times p \times (1-p) \times p$
⋮	⋮	⋮	⋮
2^{n-2}	(f,...,s,f)	1	$(1-p) \times (1-p) \ldots \times p \times (1-p)$
2^{n-1}	(f,f,...,s)	1	$(1-p) \times (1-p) \ldots \times p$
2^n	(f,f,...,f)	0	$(1-p) \times (1-p) \ldots \times (1-p)$

n independent trials, $X =$ number of successes

- ▶ Consider any outcome that results in a total of i successes.
  - ▶ This outcome will have a total of i successes and $(n - i)$ failures.
  - ▶ Probability of i success and $(n - i)$ failures $= p^i \times (1 - p)^{(n-i)}$
  .
- ▶ There number of different outcomes that result in i successes and $(n - i)$ failures$= \binom{n}{i}$
- ▶ The probability of i successes in n trials is given by

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

Section summary

▶ Independent and identically distributed distribution

▶ n independent trials

▶ Probability of i successes and $(n - i)$ failures in n independent trials.

Learning objectives

1. Derive the formula for the probability mass function for Binomial distribution.
2. Understand the effect of parameters n and p on the shape of the Binomial distribution.
3. Expectation and variance of the binomial distribution.
4. To understand situations that can be modeled as a Binomial distribution.

Binomial random variable

Definition

X is a binomial random variable with parameters n and p that represents the number of successes in n independent Bernoulli trials, when each trial is a success with probability p. X takes values $0, 1, 2, \ldots, n$ with the probability

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

Example: Tossing a coin thrice

- $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- Success= head, Failure = tails
- X is the random variable which counts the number of heads in the tosses. $n = 3$ $p = 0.5$
    - Probability mass function

X	0	1	2	3
$P(X = x_i)$	$\binom{3}{0}\frac{1}{2}^0\frac{1}{2}^3$	$\binom{3}{1}\frac{1}{2}^1\frac{1}{2}^2$	$\binom{3}{2}\frac{1}{2}^2\frac{1}{2}^1$	$\binom{3}{3}\frac{1}{2}^3\frac{1}{2}^0$
$P(X = x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Shape of the pmf for same *n* different *p*

A binomial distribution is

▶ right skewed if $p < 0.5$

▶ is symmetric if $p = 0.5$

▶ is left skewed if $p > 0.5$

. We demonstrate the same for $n = 4$ and different p

Graph of pmf of Binomial distribution- Right skewed
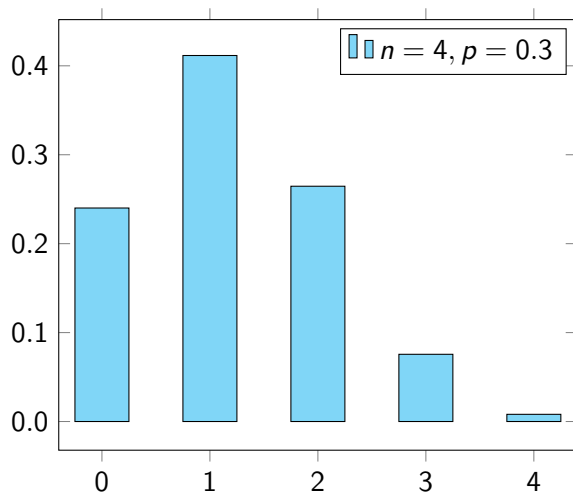
Graph of pmf of Binomial distribution- symmetric
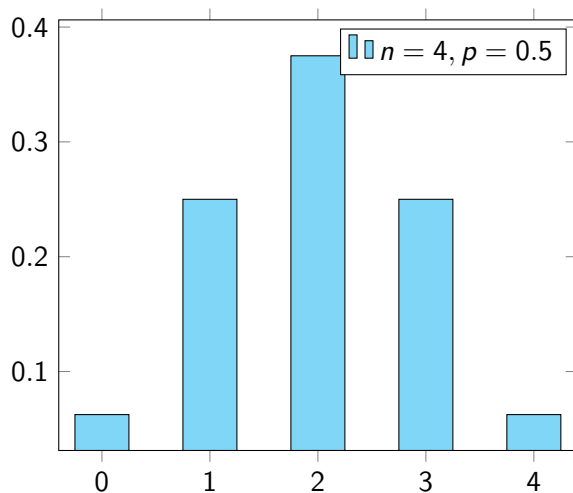
$n = 4, p = 0.8, X =$ number of successes

- Let $n = 4$ independent Bernoulli trials.
- Let $p = 0.8$ is probability of success.
- Let $X =$ number of successes in 4 independent trials.
- The probability distribution of X

X	0	1	2	3	4
$P(X = i)$	0.0016	0.0256	0.1536	0.4096	0.4096

Graph of pmf of Binomial distribution- left skewed

n = 4, p = 0.8
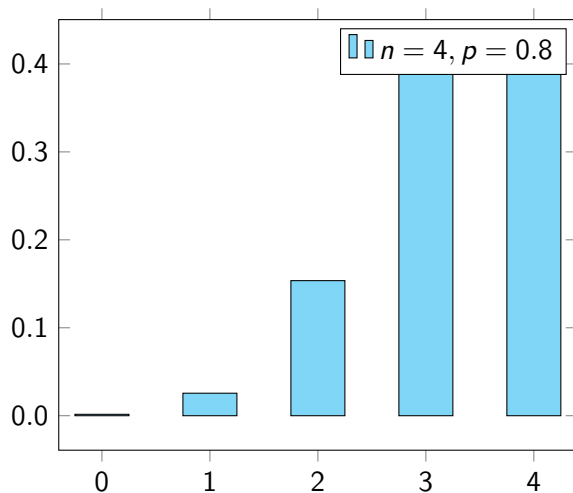
Graph of pmf of Binomial distribution- Right skewed- large n
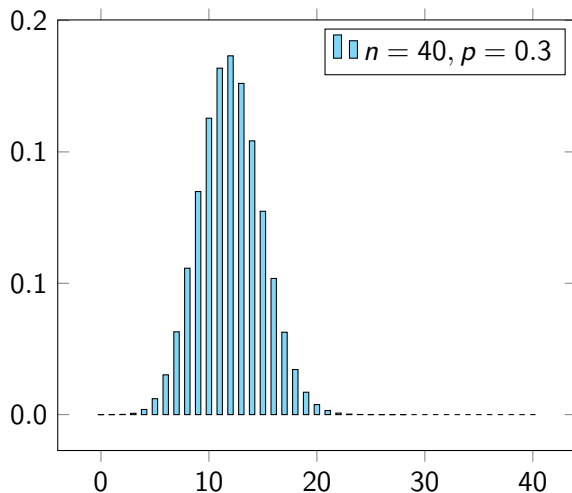
Graph of pmf of Binomial distribution- symmetric-large n
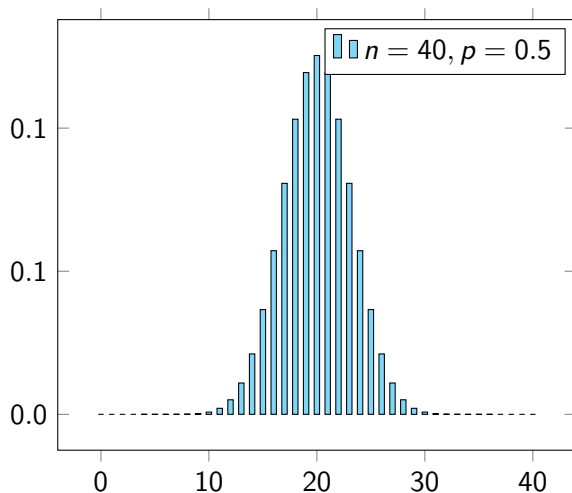
Graph of pmf of Binomial distribution- left skewed- large *n*
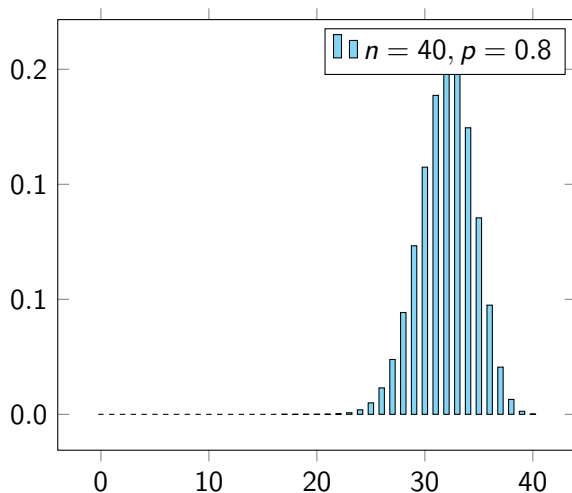
Effect of n and p on shape of distribution

- small n, small p- right skewed
- small n, large p- left skewed
- small n $p = 0.5$- symmetric
- For large n, the binomial distribution approaches symmetry.

Section summary

- ▶ Introduced the Binomial random variable and its pmf.
- ▶ Studied effect of n and p on ths shape of the distribution.

Learning objectives

1. Derive the formula for the probability mass function for Binomial distribution.
2. Understand the effect of parameters n and p on the shape of the Binomial distribution.
3. Expectation and variance of the binomial distribution.
4. To understand situations that can be modeled as a Binomial distribution.

Application: Pack of three goods

- ▶ Consider a company that sells goods in packs of three.
- ▶ The production process of the goods is not very good and results in 10% of goods being defective.
- ▶ The company believes that customers will not complain if one out of three in a pack is of bad quality, however, will complain if more than two out of three are of bad quality.
- ▶ The company wants to keep number of complaints low, say at 3%.
- ▶ How do we help the company analyse the situation?

Application: Pack of three goods

▶ Random experiment: Choosing an item and noting its quality.
$S = \{Good, Bad\}$
  ▶ Success: good
  ▶ Failure: Bad

▶ Given probability of a defective item is 0.1. Hence, Probability of good$=p=$ 0.9.

▶ We want to know number of good items in a pack of three. Hence $n = 3$

▶ Let $X =$ number of good in pack of three. X is a Binomial random variable with $n = 3, p = 0.9$.

Application: Pack of three goods-pmf

▶ The distribution of X is given by

X	0	1	2	3
$P(X = x_i)$	$\binom{3}{0}\frac{9}{10}^0\frac{1}{10}^3$	$\binom{3}{1}\frac{9}{10}^1\frac{1}{10}^2$	$\binom{3}{2}\frac{9}{10}^2\frac{1}{10}^1$	$\binom{3}{3}\frac{9}{10}^3\frac{1}{10}^0$
$P(X = x_i)$	0.001	0.027	0.243	0.729

Application: Pack of three goods-Probability of complaint

- ▶ Customers will complain if they find more than one defective in the pack of three.
  - ▶ $P(X \leq 1)$
- ▶ The distribution of X is given by

X	0	1	2	3
$P(X = x_i)$	0.001	0.027	0.243	0.729

- ▶ $P(X \leq 1) = 0.001 + 0.027 = 0.028$
- ▶ 2.8% is less than 3% which was the goal set by the company-goal achieved.
- ▶ However, if the company set 2.5% as their threshold then 2.8% would have been more than 2.5% and company would not have achieved its goal.

Effect of n- size of packs

n						
3	X	0	1	2	3	
	$P(X = i)$	0.001	0.027	0.243	0.729	
4	X	0	1	2	3	4
	$P(X = i)$	1E-04	0.0036	0.0486	0.2916	0.6561
5	X	0	1	2	3	4	5
	$P(X = i)$	1E-05	0.00045	0.0081	0.0729	0.32805	0.59049

Rolling a dice

Roll four fair dice. Define success as getting a six. Find the probability that

   a 6 appears at least once.

   b 6 appears exactly once.

   b 6 appears at least twice.

▶ Let $X =$ the number of sixes in four rolls of the dice. Then $X \sim B(4, 1/6)$.

▶ The pmf is given by

X	0	1	2	3	4
$P(X = i)$	0.4823	0.3858	0.1157	0.0154	0.0008

   a 6 appears at least once=0.5177.

   b 6 appears exactly once = 0.3858.

   c 6 appears at least twice = 0.1319.

Example: Defective ball bearings

Each ball bearing produced is independently defective with probability 0.05. If a sample of 5 is inspected, find the probability that

- a None are defective.
- b Two or more are defective.

▶ Let X = the number of defectives in sample of five. Then $X \sim B(5, 0.05)$.

▶ The pmf is given by

X	0	1	2	3	4	5
$P(X = i)$	0.7738	0.2036	0.0214	0.0011	0.0000	0.0000

- a None are defective= $P(X = 0) = 0.7738$
- b Two or more are defective=$P(X \geq 2) = 0.0225$

Example: Satellite functioning

A satellite system consists of 4 components and can function if at least 2 of them are working. If each component independently works with probability 0.8, what is the probability the system will function?

- ▶ Let $X =$ the number of components among four that are functioning. Then $X \sim B(4, 0.8)$.
- ▶ The pmf is given by

X	0	1	2	3	4
$P(X = i)$	0.0016	0.0256	0.1536	0.4096	0.4096

a System will function if $X \geq 2$, $P(X \geq 2) = 0.9728$

Example: Multiple-choice examination

A multiple-choice examination has 4 possible answers for each of 5 questions. What is the probability that a student will get 4 or more correct answers just by guessing?

a Let X be number of correct responses. $X \sim B(5, 1/4)$
The pmf is given by

X	0	1	2	3	4	5
$P(X = i)$	0.2373	0.3955	0.2637	0.0879	0.0146	0.0010

b Student will get 4 or more correct answers just by guessing
$X \geq 4, P(X \geq 4) = 0.0156$

Section summary

▶ Application of binomial model to real life examples.

Learning objectives

1. Derive the formula for the probability mass function for Binomial distribution.
2. Understand the effect of parameters n and p on the shape of the Binomial distribution.
3. Expectation and variance of the binomial distribution.
4. To understand situations that can be modeled as a Binomial distribution.

Expectation and Variance of Binomial Random Variable

- ▶ A binomial random variable $X \sim Bin(n, p)$ is equal to the number of successes in n independent trials when each trial is a success with probability p.

- ▶ We can represent X as

$$X = X_1 + X_2 + \ldots + X_n$$

  where X_i is equal to 1 if trial i is a success and is equal to 0 if trial i is a failure.

- ▶ $P(X_i = 1) = p$ and $P(X_i = 0) = (1 - p)$

Expectation and Variance of Binomial Random Variable

- $X = X_1 + X_2 + \ldots + X_n$
- $E(X) = E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$
  - $E(X) = p + p + \ldots + p = np$
- $V(X) = V(X_1 + X_2 + \ldots + X_n) = V(X_1) + V(X_2) + \ldots + V(X_n)$
  - $V(X) = p(1-p) + p(1-p) + \ldots + p(1-p) = np(1-p)$

Result

Using the fact that the expectation of the sum of random variables is equal to the sum of their expectations, we see that Expectation of a Binomial random variable X is

$$E(X) = np$$

Also, since the variance of the sum of independent random variables is equal to the sum of their variances, we have variance of a Binomial random variable is

$$Var(X) = np(1-p)$$

Example: Tossing a coin 500 times

If a fair coin is tossed 500 times, what is the standard deviation of the number of times that a head appears?

Let $X =$ the number of heads in 500 tosses of a fair coin. Then $X \sim B(500, 1/2). V(X) = 125, SD(X) = \sqrt{125} = 11.1803$

Finding probability given expectation and n

The expected number of heads in a series of 10 tosses of a coin is 6. What is the probability there are 8 heads?

Let X be number of heads. $X \sim B(10, p)$.

1. Since $E(X) = np$; $10p = 6$, hence $p = 0.6$
2. Prob there are 8 heads; $P(X = 8) = 0.121$

Find pmf given Expectation and Variance

If X is a binomial random variable with expected value 4.5 and variance 0.45, find

- a $P(X = 3)$
- b $P(X \geq 4)$

▶ $X \sim B(n, p)$

▶ $np = 4.5, np(1 - p) = 0.45$.

▶ Solving gives $n = 5$ and $p = 0.9$

- a $P(X = 3) = 0.0729$
- b $P(X \geq 4) = 0.9185$

Section summary

- ▶ Expectation and variance of Binomial random variable
- ▶ Applications

Hypergeometric distribution
    Examples of Hypergeometric distribution

Expectation and Variance of Hypergeometric distribution

Graph of pmf of the Hypergeometric distribution
    Binomial versus Hypergeometric distribution

Learning objectives

1. Derive the formula for the probability mass function for Hypergeometric distribution.
2. Expectation and variance of the Hypergeometric distribution.
3. To understand situations that can be modeled as a Hypergeometric distribution.

Limitations of Binomial distribution

▶ Suppose we are interested in finding the probability of success in *n* trials, where the trials are independent.
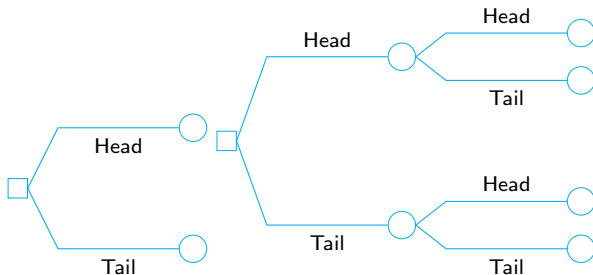
Limitations of Binomial distribution

▶ Suppose trials are not independent and the probability of
"success" is not the same for all trials.
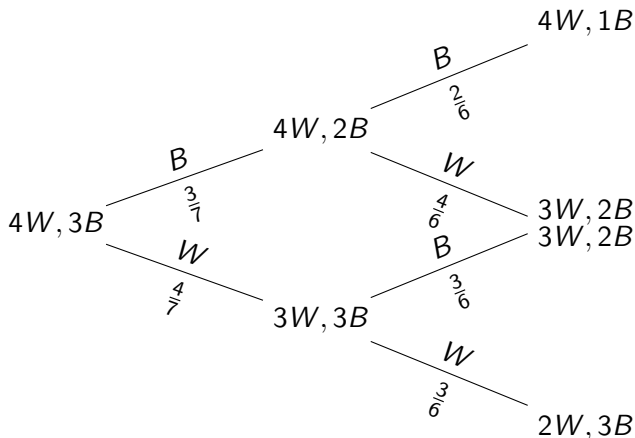
Introduction

For the hypergeometric to work,

▶ The population must be dividable into two and only two independent subsets (black balls and white balls in our example).

▶ The experiment must have changing probabilities of success with each experiment (the fact that balls are not replaced after the draw in our example makes this true in this case).

▶ Another way to say this is that you sample without replacement and therefore each draw is not independent.

The Hypergeometric distribution

▶ A discrete random variable (RV) that is characterized by:
  ▶ A fixed number of trials.
  ▶ The probability of success is not the same from trial to trial.

▶ We sample from two groups of items when we are interested in only one group.

▶ X is defined as the number of successes out of the total number of items chosen.

Understanding the Hypergeometric distribution

If we randomly select n items without replacement from a set of N items of which: m of the items are of one type and $N - m$ of the items are of a second type
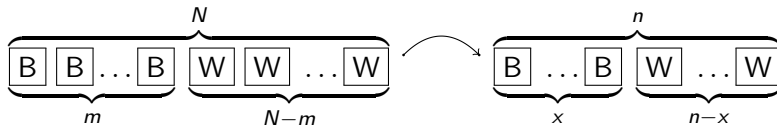
Let X be the number of items of type 1, then the probability mass function of the discrete random variable, X, is called the hypergeometric distribution and is of the form:

$$P(X = x) = \frac{\binom{m}{x}\binom{N - m}{n - x}}{\binom{N}{n}}; x = 0, 1, \ldots, n$$
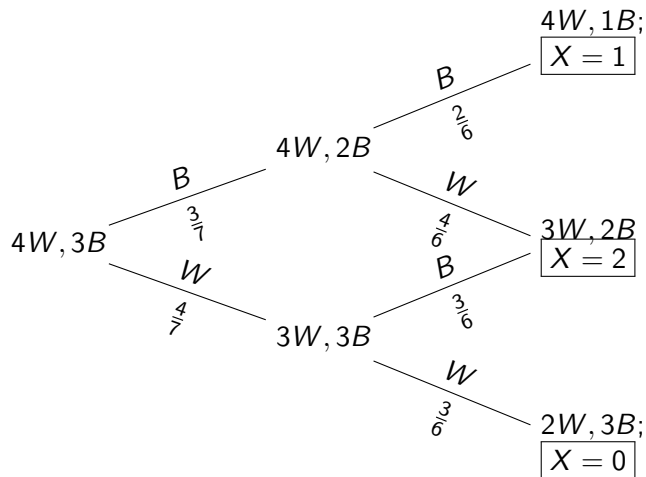
Examples: Choosing balls without replacement

A bag consists of 7 balls of which 4 are white and 3 are black. A student randomly samples two balls without replacement. Let X be the number of black balls selected.

▶ Here, $N = 7, n = 2, m = 3$

▶ X takes values: $0, 1, 2$

▶ $P(X = i) = \dfrac{\dbinom{3}{i}\dbinom{4}{2-i}}{\dbinom{7}{2}}; i = 0, 1, 2$

▶ The pmf

i	0	1	2
P(X=i)	$\frac{12}{42}$	$\frac{24}{42}$	$\frac{6}{42}$

Examples: Choosing balls without replacement

A bag consists of 50 balls of which 30 are white and 20 are blue. A student randomly samples five balls without replacement. Let X be the number of blue balls selected.

- Here, $N = 50, n = 5, m = 20$
- X takes values: $0, 1, 2, 3, 4, 5$
- $P(X = i) = \dfrac{\dbinom{20}{i}\dbinom{30}{5-i}}{\dbinom{50}{5}}; i = 0, 1, 2, 3, 4, 5$
- The pmf

i	0	1	2	3	4	5
P(X=i)	0.07	0.26	0.36	0.23	0.07	0.01

Example: Voters

▶ Assume there are 150 female voters and 250 male voters in a particular locality. If a group of twenty five voters is selected at random, then the probability that ten of the selected voters would be female can be calculated with the help of hypergeometric probability distribution.

▶ In this case: $N = 400, n = 25, m = 150$

▶

$$P(X = 10) = \frac{\binom{150}{10}\binom{250}{15}}{\binom{400}{25}}$$

Example: Defectives-1

▶ In a batch of 10 computer parts it is known that there are three defective parts. Four of the parts are selected at random to be tested. Define the random variable X to be the number of working (non defective) computer parts selected
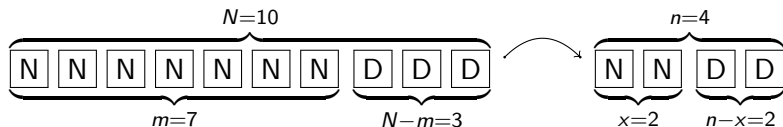
▶ This is a Hypergeometric distribution with $N = 10, n = 4, m = 7$

▶ The pmf $P(X = x) = \dfrac{\dbinom{7}{x}\dbinom{3}{4-x}}{\dbinom{10}{4}}; x = 0, 1, 2, 3, 4$

Example: Defectives-2

▶ In a batch of 10 computer parts it is known that there are three defective parts. Four of the parts are selected at random to be tested. Define the random variable X to be the number of defective computer parts selected
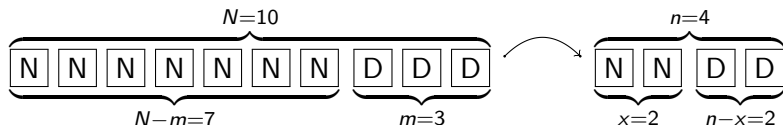
▶ This is a Hypergeometric distribution with
$N = 10, n = 4, m = 3$

▶ The pmf $P(X = x) = \dfrac{\dbinom{3}{x}\dbinom{7}{4-x}}{\dbinom{10}{4}}; x = 0, 1, 2, 3$

Example: Sampling from a deck of cards

▶ Take a deck of 52 cards. Draw five cards from the deck. Let the random variable X denote the number of aces in the random sample of five cards. What is the probability distribution of X?

▶ This is a Hypergeometric distribution with $N = 52, n = 5, m = 4$

▶ The pmf is given by

$$P(X = x) = \frac{\binom{4}{x}\binom{48}{5-x}}{\binom{52}{5}}; x = 0, 1, 2, 3, 4$$

Hypergeometric distribution

If we randomly select n items without replacement from a set of N items of which: m of the items are of one type and $N - m$ of the items are of a second type

The probability mass function of the discrete random variable is called the hypergeometric distribution and is of the form:

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}; x \le n, x \le m, n - x \le N - m$$

Section summary

- ▶ Understanding the hypergeometric distribution.
- ▶ Obtaining the probability mass function of the distribution.

Expectation

Let X follow a hypergeometric distribution in which n objects are selected from N objects with m of the objects being one type, and $N - m$ of the objects being a second type. What is the expected value of X?

$$E(X) = \frac{nm}{N}$$

Expectation- proof

$$E(X) = \sum_x x \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

Now, $\binom{m}{x} = \dfrac{m!}{x!(m-x)!}$, and, $\binom{N}{n} = \dfrac{N!}{n!(N-n)!} =$

$$\frac{N(N-1)!}{n \cdot (n-1)!(N-n)!} = \frac{N}{n} \cdot \frac{(N-1)!}{(n-1)!(N-1-(n-1))!} = \frac{N}{n} \cdot \binom{N-1}{n-1}$$

Hence, $E(X) = \sum_x x \dfrac{\dfrac{m \cdot (m-1)!}{x!(m-x)!}\binom{N-m}{n-x}}{\dfrac{N}{n} \cdot \binom{N-1}{n-1}} =$

$$\sum_x \frac{nm}{N} \frac{\dfrac{(m-1)!}{(x-1)!(m-1-(x-1))!}\binom{(N-1)-(m-1)}{(n-1)-(x-1)}}{\binom{N-1}{n-1}} = \frac{nm}{N}$$

Variance

Let X follow a hypergeometric distribution in which n objects are selected from N objects with m of the objects being one type, and $N - m$ of the objects being a second type. What is the variance of X?

$$Var(X) = n\frac{m}{N}\frac{N-m}{N}\frac{N-n}{N-1}$$

$N = 20, m = 4, n = 5$
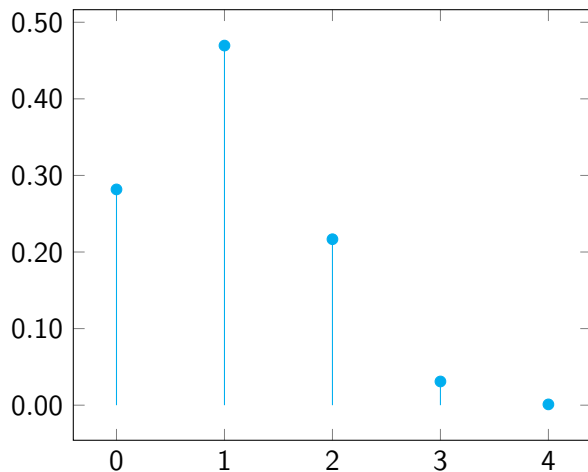
$N = 20, m = 10, n = 5$
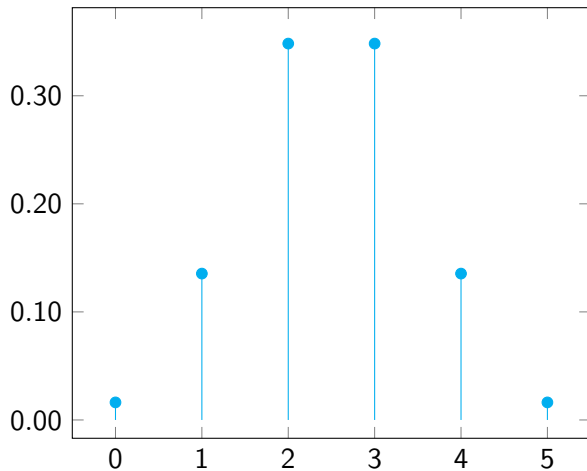
$N = 20, m = 16, n = 5$
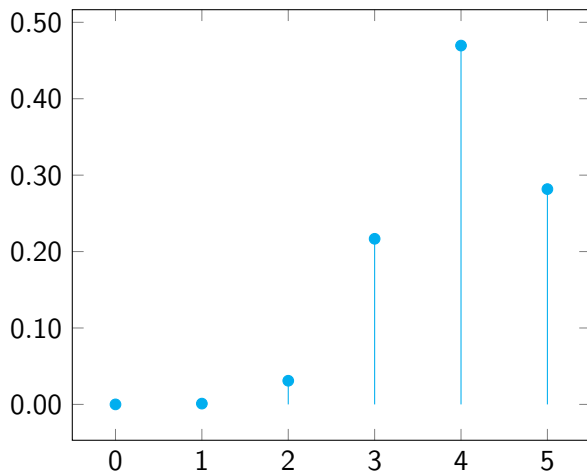
$$\bullet \quad N = 20, m = 10, n = 10$$
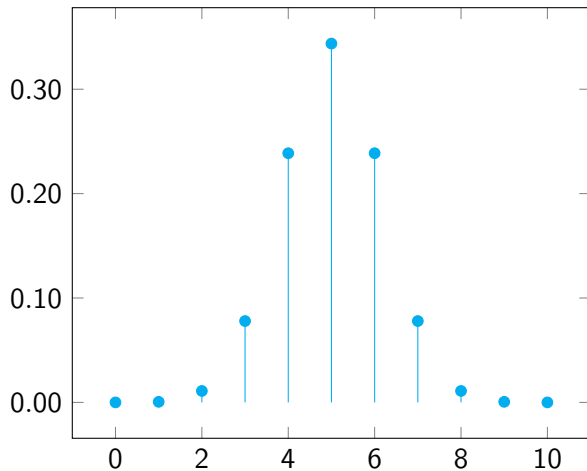
$N = 100, m = 50, n = 10$
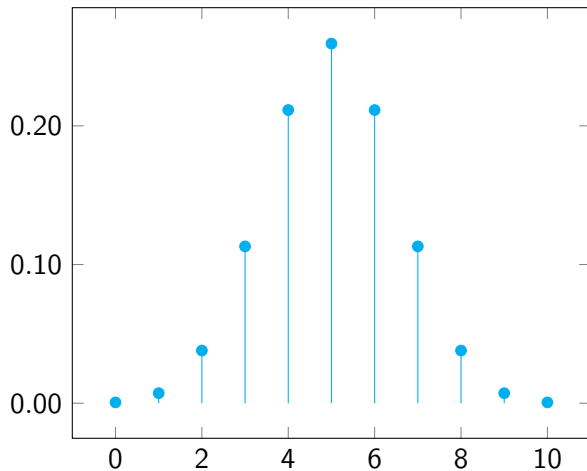
$N = 500, m = 250, n = 10$
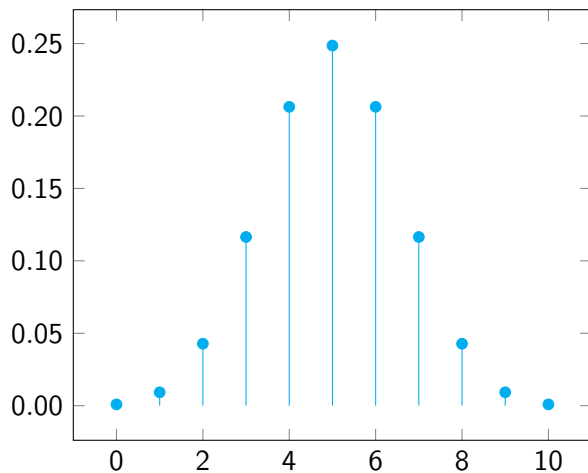
Expectation and variance

- ► $X \sim Hypergeometric(N, m, n)$
  - ► $E(X) = \dfrac{nm}{N}$
  - ► $Var(X) = n\dfrac{m}{N}\dfrac{N-m}{N}\dfrac{N-n}{N-1}$
- ► $Y \sim Bin\left(n, \dfrac{m}{N}\right)$
  - ► $E(X) = \dfrac{nm}{N}$
  - ► $Var(X) = n\dfrac{m}{N}\dfrac{N-m}{N}$
- ► $\dfrac{N-n}{N-1}$ is known as finite population correction
  - ► For $n = 1$, replacement has no effect both are Bernoulli trial
  - ► For $n = N$, the whole population is sampled- hence variance is zero.
- ► If the population N is very large compared to the sample size n (i.e. $N >> n$) then Hypergeometric(N, m, n) is about Binomial$(n, \dfrac{m}{N})$.

Binomial versus Hyergeometric distribution

$n = 5, p = 0.5$

$N = 500, m = 250, n = 5, n/N = 1\%$

Binomial versus Hyergeometric distribution

$n = 25, p = 0.5$

$N = 500, m = 250, n = 25, n/N = 5\%$

Binomial versus Hyergeometric distribution

$n = 50, p = 0.5$

$N = 500, m = 250, n = 50, n/N = 10\%$

Binomial versus Hypergeometric distribution

$n = 250, p = 0.5$

$N = 500, m = 250, n = 250, n/N = 50\%$
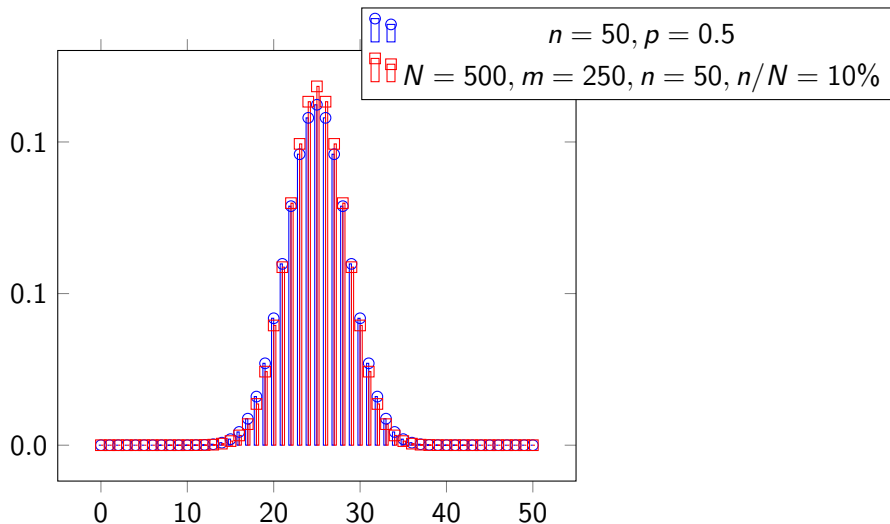
Section summary

▶ Binomial versus Hypergeometric

Learning objectives

1. Derive the formula for the probability mass function for Poisson distribution.
2. Expectation and variance of the Poisson distribution.
3. To understand situations that can be modeled as a Poisson distribution.

Introduction

▶ The Poisson probability distribution gives the probability of a number of events occurring in a fixed interval of **time** or **space**.

▶ We assume that these events happen with a known average rate, λ, and independently of the time since the last event.

▶ Let X denote the number of times an event occurs in an interval of time (or space).

▶ We say $X \sim Poisson(\lambda)$, in other words, X is a random variable that follows Poisson distribution with parameter λ.

▶ The Poisson distribution may be used to approximate the Binomial distribution if the probability of success is "small" and the number of trials is "large" .

Motivation example

Consider a researcher who is observing the number of vehicles that pass a busy traffic intersection in a day. She collects data comprising of 1000 one minute intervals and tabulates the same in form of a frequency table given below.

Number of vehicles	0	1	2	3	4	5	> 6
Count	80	225	260	225	110	70	30
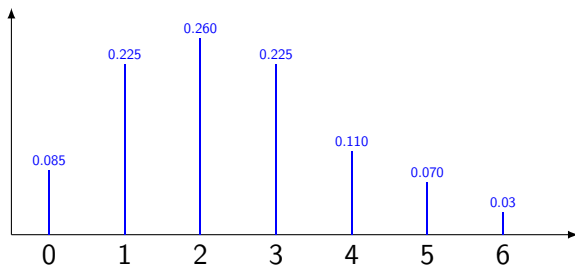
Tabular summary

x	Freq f	Rel Freq f_r	$f_r x$	$f_r x^2$
0	80	0.08	0	0
1	225	0.225	0.225	0.225
2	260	0.26	0.52	1.04
3	225	0.225	0.675	2.025
4	110	0.11	0.44	1.76
5	70	0.07	0.35	1.75
6	30	0.03	0.18	1.08
	1000	1	2.39	7.88

▶ Mean= 2.39

▶ Variance $= 7.88 - 2.39^2 = 2.16$

Observations

- Number of vehicles passing a traffic intersection are at random and independently of each other
- The average number of vehicles per minute is about 2.39 which is equivalent to 143 per hour.
- Question: What is the appropriate probability distribution to model the number of vehicles passing a traffic intersection?
  - Poisson

Derivation

Let X denote the number of events in a given interval (time or space). Then X follows a Poisson distribution with parameter λ

1. The number of events occurring in non-overlapping intervals are independent.

2. The probability of exactly one event in a short interval of length, δt, is equal to $\lambda \delta t$.

3. The probability of exactly two or more events in a short interval is essentially zero.

What is the Probability of n events happening in interval of length t?

Poisson as Binomial approximation

▶ Define "success" as exactly one event happening in a short interval of length δt

▶ The n events happening in interval of length t can be viewed as n successes happening in n intervals of length δt, with each one of them being an independent and identical trial.

▶ Hence the problem can be viewed as a $Bin\left(n, p = \dfrac{\lambda}{n}\right)$ experiment.

Derivation- contd

$$
\begin{aligned}
&= && \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= && \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= && \frac{\lambda^x}{x!} \left(\frac{n(n-1)\dots(n-x+1)}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= && \frac{\lambda^x}{x!} \left(\frac{n^x(1 - \frac{1}{n})\dots(1 - \frac{(x-1)}{n})}{n^x}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x}
\end{aligned}
$$

Now let's make the intervals very small, i.e, $\delta t \longrightarrow 0$ or $n \longrightarrow \infty$

Derivation- contd

$$
\begin{aligned}
&= \lim_{n \to \infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n \to \infty} \frac{\lambda^x}{x!} \underbrace{\left(\frac{n^x(1 - \frac{1}{n}) \ldots (1 - \frac{(x-1)}{n})}{n^x}\right)}_{\to 1 \ as \ n \to \infty} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda} \ as \ n \to \infty} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\to 1 \ as \ n \to \infty} \\
&= \frac{\lambda^x}{x!} e^{-\lambda}
\end{aligned}
$$

Probability mass function of Poisson

The distribution, with an average number of λ events per interval, is defined as Poisson discrete random variable, $X \sim Poisson(\lambda)$, with the p.m.f given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$$

- ▶ X represents the random variable number of events per time interval (In the example: number of vehicles passing per minute)
- ▶ e is the mathematical constant 2.718

Going back to the example

x	freq	Prob
0	80	0.092
1	225	0.219
2	260	0.262
3	225	0.208
4	110	0.125
5	70	0.060
6	30	0.024
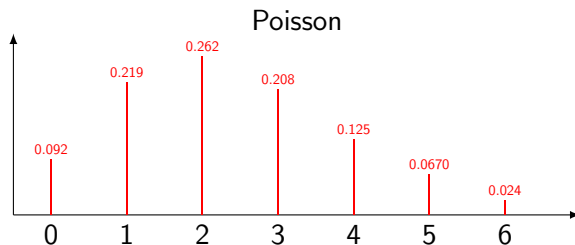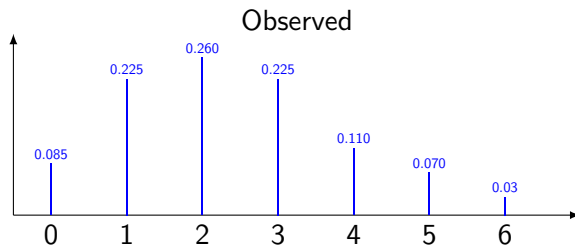
Shape of pmf versus λ

▶ The shape of the Poisson distribution depends on the value of the parameter λ.

▶ If λ is small the distribution has positive skew, but as λ increases the distribution becomes progressively more symmetrical.

Graph of pmf for $\lambda = 1$
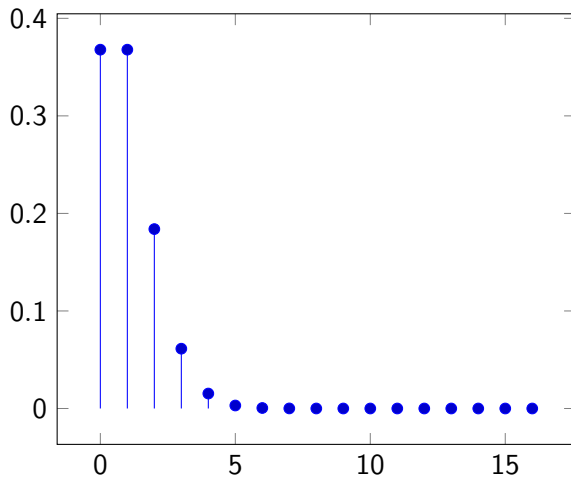
Graph of pmf for $\lambda = 4$
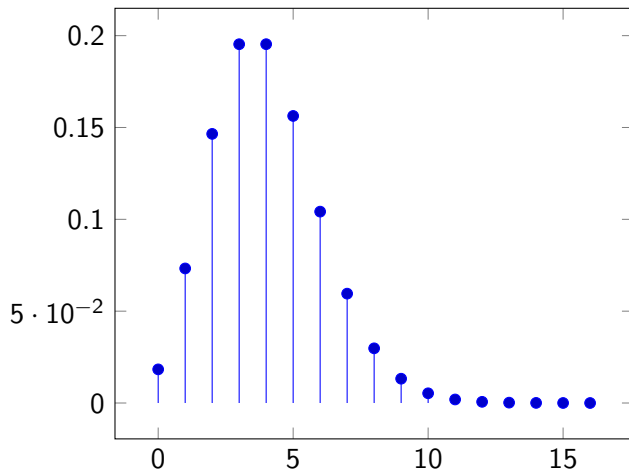
Graph of pmf for $\lambda = 8$
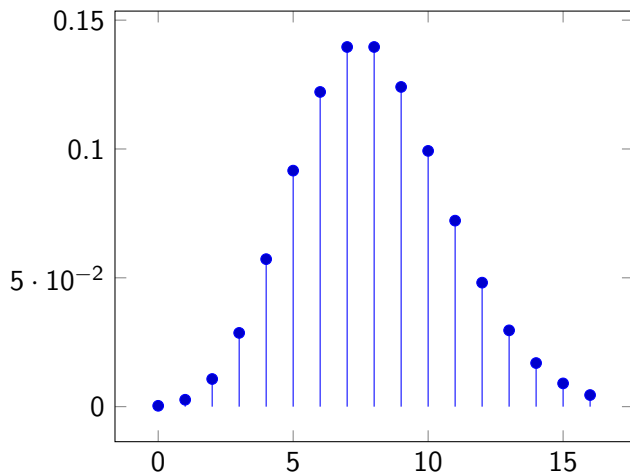
Section summary

- pmf of Poisson distribution
- shape of pmf versus λ

Expectation of Poisson distribution

$$
\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda}{x} \frac{\lambda^{(x-1)}}{(x-1)!} \\
&= e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{(x-1)}}{(x-1)!} \\
&= e^{-\lambda}\lambda \left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \dots \right) \\
&= e^{-\lambda}\lambda e^{\lambda} \\
&= \lambda
\end{aligned}
$$

Variance of Poisson distribution

$$
\begin{aligned}
E(X(X-1)) &= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^2}{x(x-1)} \frac{\lambda^{(x-2)}}{(x-2)!} \\
&= e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{(x-2)}}{(x-2)!} \\
&= e^{-\lambda} \lambda^2 \left(\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \ldots \right) \\
&= e^{-\lambda} \lambda^2 e^{\lambda} \\
&= \lambda^2
\end{aligned}
$$

Variance of Poisson distribution

▶ Now, $E(X^2) = E(X(X-1)) + E(X) = \lambda^2 + \lambda$.

▶ Hence

$$Var(X) = E(X^2) - (E(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

▶ For a Poisson random variable $X \sim Poisson(\lambda)$, both the expected value and the variance of X are equal to λ.

Statistics for Data Science -1
└─ Applications of Poisson distribution
  └─ Modeling in time

Examples of Poisson distribution

► Events occurring in fixed interval of time
1. Number of vehicles passing through a traffic intersection in a fixed time interval of one minute.
2. Number of people withdrawing money from a bank in a fixed time interval of fifteen minutes.
3. Number of telephone calls received per minute at a call center

► Events occurring in fixed interval of space
1. Number of typos (incorrect spelling) in a book.
2. Number of defects in a wire cable of finite length.
3. Number of defects per meter in a roll of cloth

Statistics for Data Science -1
└─Applications of Poisson distribution
  └─Modeling in time

Modeling number of accidents

Suppose the number of accidents per week in a factory can be modeled by the Poisson distribution with a mean of 0.5.

1. Find the probability that in a particular week there will be less than two accidents?
   Let X be the number of accidents per week in the factory.
   We have $X \sim Poisson(\lambda = 0.5)$
   Need to find $P(X \leq 2) = \sum_{i=0}^{2} \frac{e^{-0.5} \times 0.5^{i}}{i!} =$
   $= 0.6065 + 0.3033 + 0.0758 = 0.9856$

Modeling number of killings

The number of dogs that are killed on a particular stretch of road in Chennai in any one day can be modeled by a Poisson(0.42) random variable.

1. Calculate the probability that exactly two dogs are killed on a given day on this stretch of road.

   Let $X=$ number of dogs killed in one day

   $X \sim Poisson(\lambda = 0.42)$

   $P(X=2) = \dfrac{e^{-0.42} \times 0.42^2}{2!} = 0.058$

2. Find the probability that exactly two dogs are killed over a 5-day period on this stretch of road.

   Let $X=$ number of dogs killed in five day

   $X \sim Poisson(\lambda = 2.1)$

   $P(X=2) = \dfrac{e^{-2.1} \times 2.1^2}{2!} = 0.27$

Statistics for Data Science -1
└─ Applications of Poisson distribution
   └─ Modeling in time

Modeling the number of defects

Suppose the number of defects in a wire cable can be modeled by the Poisson distribution with a of 0.5 defects per meter.

1. Find the probability that a single meter of wire will have exactly 3 defects
   Let X be the number of defects per meter
   We have $X \sim Poisson(\lambda = 0,5)$
   Need to find $P(X = 3) = \dfrac{e^{-3} \times 0.5^3}{3!} = 0.0126$

Modeling typos

A typist makes 1500 mistakes in a book of 500 pages. Let X be number of mistakes per page. Then, $X \sim Poisson(3)$

On how many pages would you expect to find

1. no mistake

   $P(X = 0) = \dfrac{e^{-3} \times 3^0}{0!} = 0.0498$

   Number of pages with no mistakes $= 0.0498 \times 500 \approx 25$ pages

2. one mistake

   $P(X = 1) = \dfrac{e^{-3} \times 3^1}{1!} = 0.1494$

   Number of pages with no mistakes $= 0.1494 \times 500 \approx 75$ pages

3. three or more mistakes

   $P(X >= 31) = 1 - \dfrac{e^{-3} \times 3^0}{0!} - \dfrac{e^{-3} \times 3^1}{1!} - \dfrac{e^{-3} \times 3^2}{2!} = 0.5768$

   Number of pages with no mistakes $= 0.5768 \times 500 \approx 288$ pages

Section summary

Modeling situations using Poisson distribution

- ▶ in time
- ▶ in space

Learning objectives

1. Define what is a continuous random variable.
2. Probability distribution function and examples
3. Cumulative distribution function, graphs, and examples.
4. Expectation and variance of random variables.

Probability density function, graph, and examples
  Probability density function

Discrete and Continuous random variables

Definition
A random variable that can take on at most a countable number of possible values is said to be a discrete random variable.

Definition
When outcomes for random event are numerical, but cannot be counted and are infinitely divisible, we have continuous random variables.

Discrete and continuous random variable

▶ A discrete random variable is one that has possible values that are discrete points along the real number line.

▶ A continuous random variable is one that has possible values that form an interval along the real number line. In other words, a continuous random variable can assume any value over an interval or intervals.

Probability density function (pdf)

- ▶ Every continuous random variable X has a curve associated with it.
- ▶ The probability distribution curve of a continuous random variable is also called its probability density function. It is denoted by $f(x)$

Area under a pdf

▶ Consider any two points a and b, where a is less than b.

▶ The probability that X assumes a value that lies between a and b is equal to the area under the curve between a and b. That is,

$P(X \in [a, b]) = P(a \leq X \leq b)$ is area under curve between a and b

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

Properties of pdf

1. The area under the probability distribution curve of a continuous random variable between any two points is between 0 and 1.

Properties of pdf

2. Total area under the probability distribution curve of a continuous random variable is always 1.

Properties of pdf

▶ The area under the graph of the probability density function between points a and b is the same regardless of whether the endpoints a and b are themselves included:

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

▶ The probability density curve of a random variable X is a curve that never goes below the $x-$ axis

Example

Figure below is a probability density function for the random variable that represents the time (in minutes) it takes a repairer to service a television. The numbers in the regions represent the areas of those regions.

What is the probability that the repairer takes

1. Less than 20 =0.29

2. Less than 40 =0.56

3. More than 50 =0.33

4. Between 40 and 70 minutes to complete a repair? =0.27

Cumulative distribution function

For a continuous random variable X

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

Since the probability that a continuous random variable X assumes a single value is always zero, we have

$$P(X < a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

Expectation and Variance

- Expected value: $E(X) = \int x\, f(x)dx$.
- Variance: $Var(X) = \int (x - E(X))^2\, f(x)dx$

Section summary

- ▶ Probability density function and its properties.
- ▶ cdf, expectation, and variance of continuous random variables.

Introduction

▶ A random variable is said to be uniformly distributed over the interval $[0, 1]$ if its probability density function is given by

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Learning objectives

1. Define what is a continuous random variable.
2. Probability distribution function and examples
3. Cumulative distribution function, graphs, and examples.
4. Expectation and variance of random variables.

Uniform distribution-pdf
  Standard uniform distribution

cdf of Uniform distribution

Expectation and variance of Uniform distribution

Uniform distribution $U(a, b)$

▶ A continuous random variable has a uniform distribution, denoted $X \sim U(a, b)$,
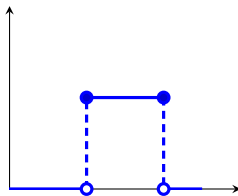
probability density function is:

$$f(x) = \begin{cases} \frac{1}{(b-a)} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Graph of pdf of a Uniform distribution $U(2, 4)$
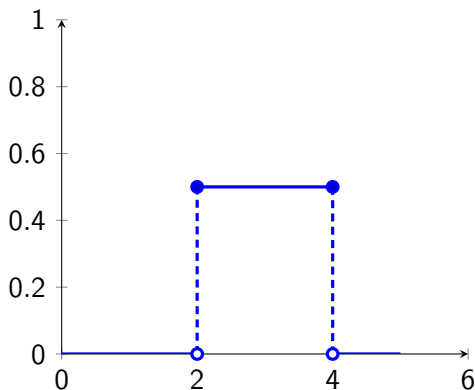
Graph of pdf of a Uniform distribution $U(-2, 2)$
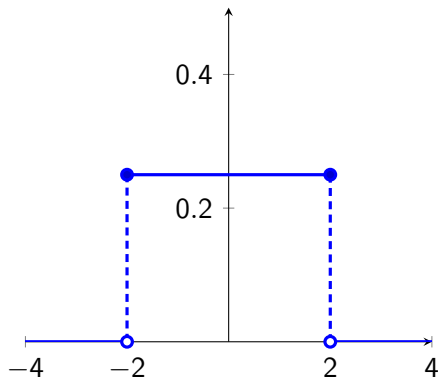
Standard uniform distribution

▶ A random variable has the standard uniform distribution with minimum 0 and maximum 1 if its probability density function is given by

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

The standard uniform distribution plays an important role in random variate generation.

▶ Verify $f(x)$ is a pdf
  ▶ $f(x) \geq 0$, for $0 < x < 1$
  ▶ $\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} f(x)dx = 1$

Statistics for Data Science -1
└─ Uniform distribution-pdf
  └─ Standard uniform distribution

Graph of pdf of a Standard uniform distribution $U(0,1)$
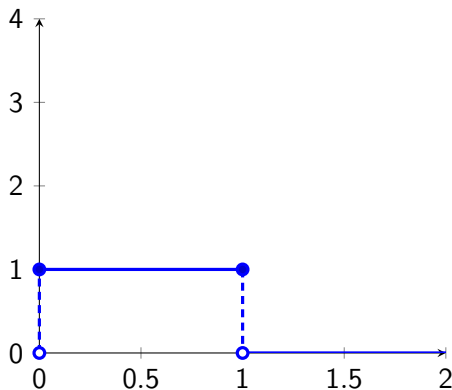
Cumulative distribution of Uniform distribution

For $X \sim U(a, b)$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b) \\ 1 & \text{for } x \geq b \end{cases}$$

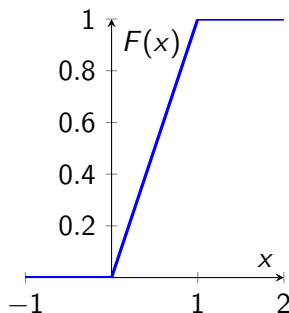# Cumulative distribution of Uniform distribution
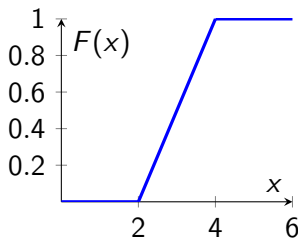
- When $x < a$ $F(x) = 0$
- When $x > b$ F(x) =1
- $a < x < b$ The slope of the line between and is $\frac{1}{(b-a)}$.

Cumulative distribution of standard uniform distribution

For $X \sim U(0,1)$

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \in [0,1) \\ 1 & \text{for } x \geq 1 \end{cases}$$

Expectation of $X \sim U(a, b)$

▶ $X \sim U(a, b)$;

$$E(X) = \frac{a + b}{2}$$

▶

$$
\begin{aligned}
E(X) &= \int_a^b x f(x) \, dx \\[2mm]
&= \int_a^b x \frac{1}{b-a} dx \\[2mm]
&= \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b \\[2mm]
&= \frac{b+a}{2}
\end{aligned}
$$

Variance of $X \sim U(a, b)$

▶ $X \sim U(a, b)$;

$$Var(X) = \frac{(b - a)^2}{12}$$

▶

$$
\begin{aligned}
E(X^2) &= \int_a^b x^2 f(x), dx &= \int_a^b x^2 \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \frac{x^3}{3} \Big|_a^b &= \frac{b^2 + a^2 + ab}{3}
\end{aligned}
$$

▶

$$
\begin{aligned}
Var(X) &= E(X^2) - [E(X)]^2 \\
&= \frac{b^2 + a^2 + ab}{3} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{(b - a^2)}{12}
\end{aligned}
$$

Example: Computing probabilities given distribution

Suppose that X is a uniform random variable over the interval $(0, 1)$. Find

1. $P(X > 1/3) = 2/3$
2. $P(X \leq 0.7) = 0.7$
3. $P(0.3 < X \leq 0.9) = 0.6$
4. $P(0.2 \leq X < 0.8) = 0.6$

Example: Application-solution

Let X denote the amount of time you will have to wait.
$X \sim U(0, 60)$

1. At least 30 minutes $= P(X \geq 30) = 30/60 = 1/2$
2. Less than 15 minutes $= P(X < 15) = 15/60 = 1/4$
3. Between 10 and 35 minutes $= P(10 \leq X \leq 35) = 25/60 = 5/12$
4. Less than 45 minutes $= P(X < 45) = 45/60 = 3/4$

Section summary

- ▶ Uniform distribution
- ▶ Standard uniform distribution
- ▶ applications

Learning objectives

1. Define what is a continuous random variable.
2. Probability distribution function and examples
3. Cumulative distribution function, graphs, and examples.
4. Expectation and variance of random variables.

Non Uniform distribution

Triangular distribution

Example

Suppose that the number of minutes of playing time of a certain college basketball player in a randomly chosen game has the following density curve.
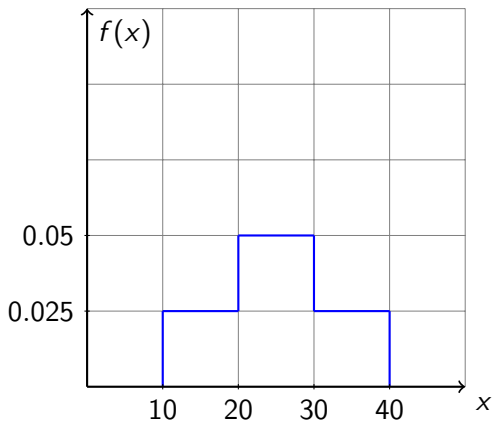
Questions

Find the probability that the player plays

1. Over 20 minutes
2. Less than 25 minutes
3. Between 15 and 35 minutes
4. More than 35 minutes

Solution

Let X be amount of playing time in minutes. Find the probability that the player plays

1. Over 20 minutes $= P(X > 20) = 0.5 + 0.25 = 0.75$
2. Less than 25 minutes $= P(X < 25) = 0.25 + 0.25 = 0.5$
3. Between 15 and 35 minutes $= P(15 \leq X \leq 35) = 0.125 + 0.5 + 0.125 = 0.75$
4. More than 35 minutes $= P(X > 35) = 0.125$

Triangular distribution

It is now 2 p.m., and Joan is planning on studying for her statistics test until 6 p.m., when she will have to go out to dinner. However, she knows that she will probably have interruptions and thinks that the amount of time she will actually spend studying in the next 4 hours is a random variable whose probability density curve is as follows:
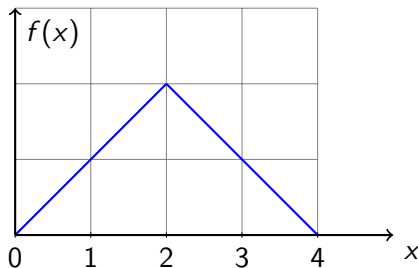
Questions

1. What is the height of the curve at the value 2?
2. What is the probability she will study more than 3 hrs?
3. What is the probability she will study between 1 and 3 hrs?

Solution
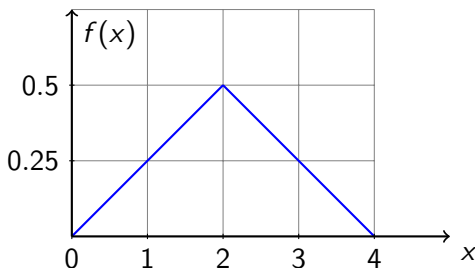
1. What is the height of the curve at the value 2?$=1/2 unit$
2. What is the probability she will study more than 3 hrs?$=1/8$
3. What is the probability she will study between 1 and 3 hrs?$=3/4$

Section summary

- ▶ Non uniform distribution
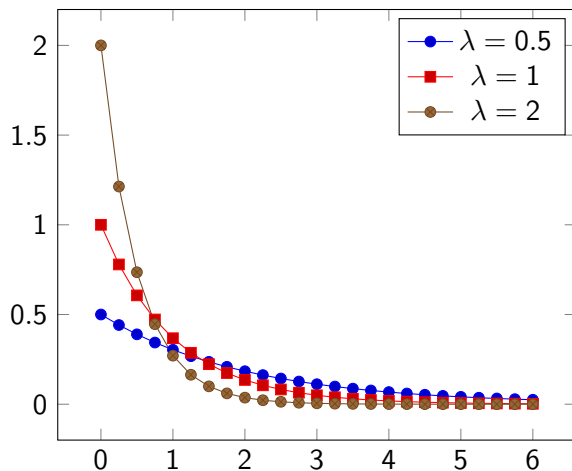- ▶ Triangular distribution

Exponential distribution

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

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

is said to be an exponential random variable (or, more simply, is said to be exponentially distributed) with parameter λ.
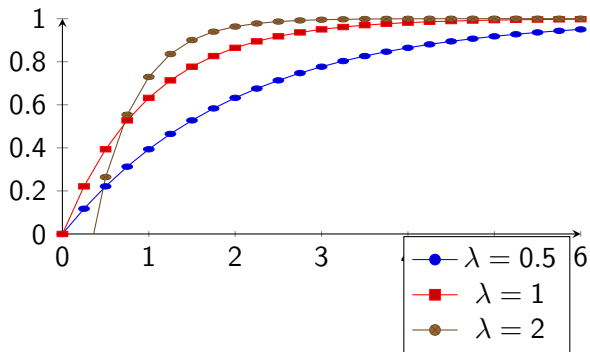
Graph of pdf for different values of λ

cdf of Exponential distribution

$$
\begin{aligned}
F(a) \ &= P(X \leq a) \\
&= \int_0^a \lambda e^{-\lambda x} dx \\
&= -e^{-\lambda x} \Big|_0^a \\
&= 1 - e^{-\lambda a}
\end{aligned}
$$

Graph of cdf for different values of λ

Expectation and variance of exponential distribution

$X \sim exp(\lambda)$

- $E(X) = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$

- It can be shown through integration by parts

$$E(X^n) = \frac{n}{\lambda} E(X^{n-1})$$

- $E(X) = \frac{1}{\lambda}$
- $E(X^2) = \frac{2}{\lambda} \frac{1}{\lambda} = \frac{2}{\lambda^2}$
- Hence $Var(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$

Thus, the mean of the exponential is the reciprocal of its parameter λ, and the variance is the mean squared.

Application

In practice, the exponential distribution often arises as the distribution of the amount of time until some specific event occurs.

- ▶ Suppose that the length of a phone call in minutes is an exponential random variable with parameter $\lambda = 0.1$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait
  - a more than 10 minutes
  - b between 10 and 20 minutes.

Solution Let X denote the length of the call made by the person in the booth. $X \sim exp(0.5)$

- a more than 10 minutes $= P(X > 10) = e^{-1} \approx 0.368$

- b between 10 and 20 minutes$=$
  $P(10 < X < 20) = F(20) - F(10) = e^{-1} - e^{-2} \approx 0.233$

Section summary

▶ Exponential distribution and its applications.