

**UNIVERSITY OF RWANDA**

**COLLEGE OF SCIENCE AND TECHNOLOGY**

**SCHOOL OF ICT**

**DEPARTMENT OF IT**

**ASSIGNEMENT: STATISTICAL METHODS OF DATA ANALYSIS**

**GROUP 3 MEMBERS:**

|  |  |
| --- | --- |
| NAMES | REGISTRATION NUMBERS |
| 1.ARERERWENEZA Charite | 217094236 |
| 2.BYIRINGIRO Yves | 217087558 |
| 3.HAKIZIMANA Jackson | 217140211 |
| 4.KARASANYI Johnson | 217034853 |
| 5.NDANYUZWE Gentil | 217015689 |
| 6. DUSHIMIYIMANA MUH0ZA PACIFIQUE | 217174817 |
| 7. UMUHOZA CLARISSE | 217058256 |

Table of Contents

[Chapter1: Introduction to data 4](#_Toc26969349)

[1.1 what is case study? 4](#_Toc26969350)

[1.2 Data basics 4](#_Toc26969351)

[1.2.1 Types of variables 4](#_Toc26969352)

[1.3 Overview of data collection principles 6](#_Toc26969353)

[1.3.1 Populations and samples 6](#_Toc26969354)

[1.3.2 Anecdotal evidence 7](#_Toc26969355)

[1.3.3 Population vs Sample 7](#_Toc26969356)

[1.3.4 Sampling bias 8](#_Toc26969357)

[1.3.5 Explanatory variable 8](#_Toc26969358)

[1.4 Observational studies and experiments 8](#_Toc26969359)

[1.4.1 Prospective vs. retrospective studies 9](#_Toc26969360)

[1.4.2 Obtaining good samples 9](#_Toc26969361)

[1.4.2.1 Stratified sample 10](#_Toc26969362)

[1.4.2.2 Cluster sample 10](#_Toc26969363)

[1.5 Principles of experimental design 11](#_Toc26969364)

[1.5.1. Difference between blocking and explanatory 13](#_Toc26969365)

[1.6 Examining numerical data 15](#_Toc26969366)

[1.6.1 Scatterplots 15](#_Toc26969367)

[**Scatterplots** are useful for visualizing the relationship between two numerical variables. 15](#_Toc26969368)

[1.6.2 Dot plots & mean 15](#_Toc26969369)

[1.6.3 Histograms 16](#_Toc26969370)

[1.6.4 Variance and standard deviation 17](#_Toc26969371)

[1.6.5 Intensity maps 17](#_Toc26969372)

[1.7 Considering categorical data 18](#_Toc26969373)

[1.7.1 Contingency tables 18](#_Toc26969374)

[1.7.2 Bar plots 19](#_Toc26969375)

[1.7.3 Pie charts 20](#_Toc26969376)

[Chapter2: Probability 20](#_Toc26969377)

[2.1 probability 20](#_Toc26969378)

[2.1.1 Disjoint or mutually exclusive outcomes 21](#_Toc26969379)

[2.1.2 Probabilities when events are not disjoint 21](#_Toc26969380)

[2.1.3 Probability distributions 21](#_Toc26969381)

[2.1.4 Complement of an event 22](#_Toc26969382)

[2.1.5 Independence 22](#_Toc26969383)

[2.2 Conditional probability 22](#_Toc26969384)

[2.2.1 Marginal and joint probabilities 23](#_Toc26969385)

[2.2.2 General multiplication rule 24](#_Toc26969386)

[2.2.3 Independence considerations in conditional probability 25](#_Toc26969387)

[2.2.4 Tree diagrams 25](#_Toc26969388)

[2.2.5 Bayes’ Theorem 25](#_Toc26969389)

[2.3 Random variables 27](#_Toc26969390)

[2.3.1 Variability in random variables 27](#_Toc26969391)

[2.3.2 Linear combinations of random variables 28](#_Toc26969392)

[2.3.3 Variability in linear combinations of random variables 28](#_Toc26969393)

[2.3.4 Variability of linear combinations of random variables 29](#_Toc26969394)

# **Chapter1: Introduction to data**

## **1.1 what is case study?**

The term case study has multiple meanings.it can be used to describe a detailed study of a single social unit or to describe a research method.

Case study refers to the process or record of research into the development of a particular person, group, or situation over a period of time.

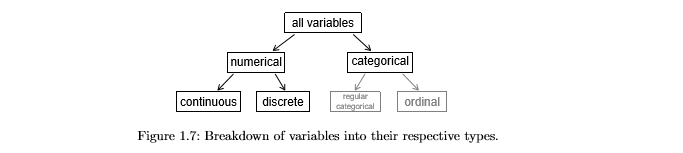
Case study “investigate a contemporary phenomenon within its real-life context when the boundaries between phenomena and context are not clearly evident and multiple sources of evidence are used.

The results of case study cannot be generalizable.

We cannot immediately generalize the results of case study sample, but it is encouraging.

# **1.2 Data basics**

## **1.2.1 Types of variables**



Variables are in two types;

**- Categorical**

**- Numerical**

A **numerical variable** is similar to an ordinal variable, except that the intervals between the values of the numerical variable are equally spaced.

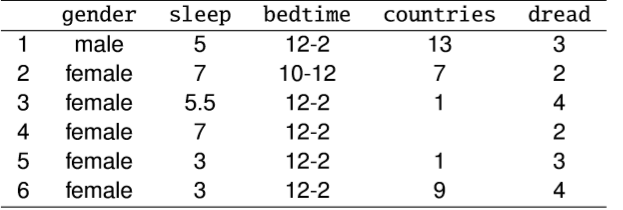
**Numerical is classified into two:**

* A **continuous variable** is a variable whose value is obtained by measuring
* A **discrete variable** is a variable whose value is obtained by counting.

**Categorical is classified into two:**

* **A categorical variable** (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories. For example, gender is a categorical variable having two categories (male and female) and there is no intrinsic ordering to the categories.
* **An ordinal variable** is similar to a categorical variable. The difference between the two is that there is a clear ordering of the variables. For example, suppose you have a variable, economic status, with three categories (low, medium and high).

In addition to being able to classify people into these categories, you can order the categories as low, medium and high



**gender:** categorical

**sleep:** numerical, continuous

**bedtime**: categorical, ordinal

**countries:** numerical, discrete

**dread:** categorical, ordinal - could also be used as numerical

## **1.3 Overview of data collection principles**

* Populations and samples
* Anecdotal evidence
* evidence based on a limited sample size that might not be representative of the population
* Sampling from a population
* Explanatory and response variables
* Observational studies and experiments

### **1.3.1 Populations and samples**

Consider the following research question

1. What is the average mercury content in swordfish in the Atlantic Ocean?

The target population is all swordfish in the Atlantic Ocean, and each swordfish represents a case. Often

Times, it is too expensive to collect data for every case in a population. Instead, a sample

is taken. A sample represents a subset of the cases and is often a small fraction of the

population. For instance, 60 swordfish (or some other number) in the population might

be selected, and this sample data may be used to provide an estimate of the population

average and answer the research question.

### **1.3.2 Anecdotal evidence**

Consider the following research question

1. A man on the news got mercury poisoning from eating swordfish, so the average

mercury concentration in swordfish must be dangerously high.

Each of the conclusions are based on some data. However, there are two problems. First,

the data only represent one or two cases. Second, and more importantly, it is unclear

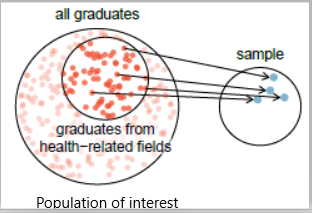
whether these cases are actually representative of the population. Data collected in this

haphazard fashion is called anecdotal evidence.

Anecdotal evidence

* Be careful of data collected in a haphazard fashion. Such evidence may be true
* and verifiable, but it may only represent extraordinary cases.

### **1.3.3 Population vs Sample**



* Figure 1.12: Instead of sampling from all graduates equally, a nutrition
* major might inadvertently pick graduates with health-related majors dis-proportionately often.

### **1.3.4 Sampling bias**

**Non-response:** If only a small fraction of the randomly sampled people chooses to respond to a survey, the sample may no longer be representative of the population.

**Voluntary response:** Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the

population.

**Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample.

### **1.3.5 Explanatory variable**

To identify the explanatory variable in a pair of variables, identify

which of the two is suspected of affecting the other:

explanatory variable **might affect** response variable.

## **1.4 Observational studies and experiments**

**Observational study:** Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely “observe”, and can only establish an association between the explanatory and response variables.

**Experiment:** Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.

### **1.4.1 Prospective vs. retrospective studies**

**A prospective study** identifies individuals and collects information as events unfold.

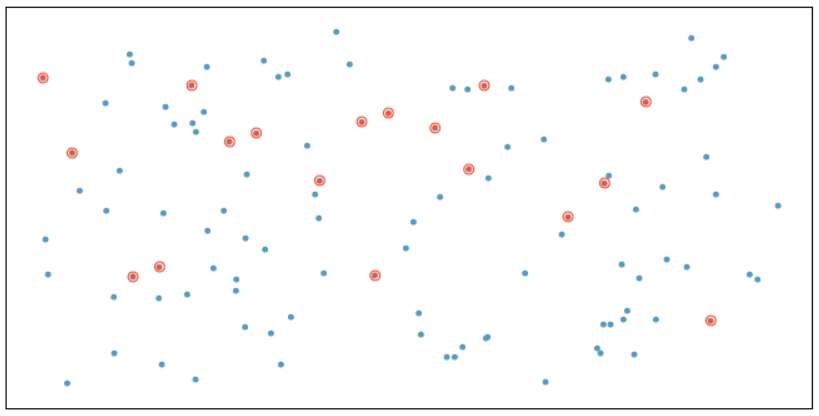
Example: The Nurses’ Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.

**Retrospective studies** collect data after events have taken place.

Example: Researchers reviewing past events in medical records.

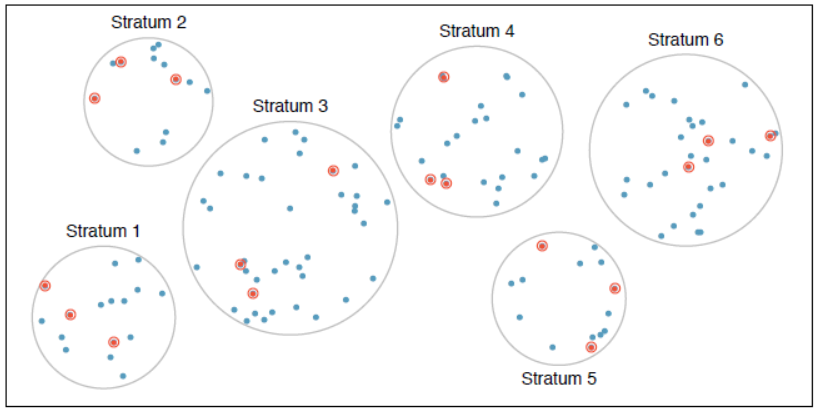
### **1.4.2 Obtaining good samples**

* Most commonly used random sampling techniques are **simple**, **stratified, and cluster sampling**.
* **Simple random sample:** Randomly select cases from the population, where there is no implied connection between the points that are selected.



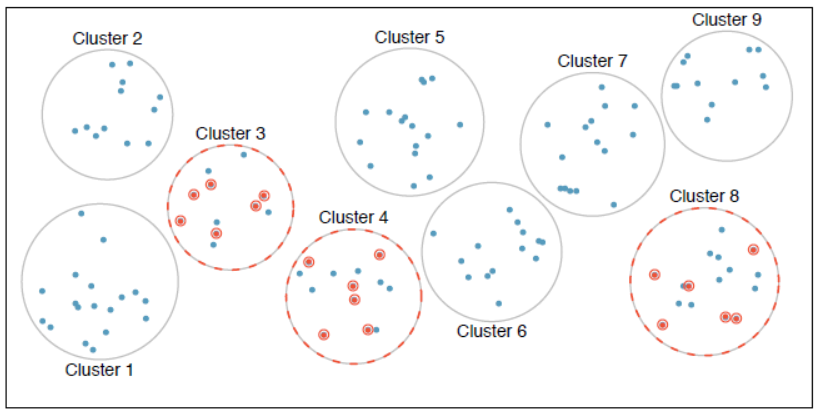
### **1.4.2.1 Stratified sample**

Strata are made up of similar observations. We take a simple random sample from each stratum.



### **1.4.2.2 Cluster sample**

Clusters are usually not made up of homogeneous observations, and we take a simple random sample from a random sample of clusters. Usually preferred for economic reasons.

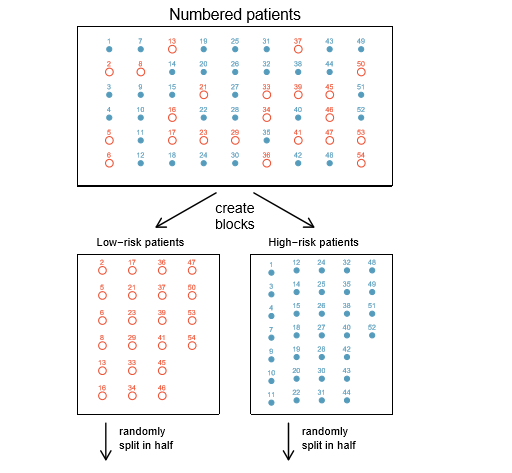


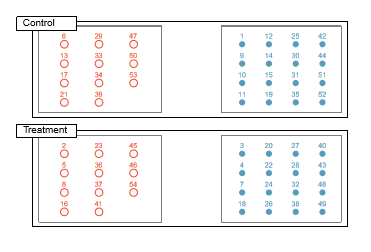
## 1.5 Principles of experimental design

* **Control:** Compare treatment of interest to a control group.
* **Randomize:** Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
* **Replicate:** Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
* **Block:** If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.

For instance, if we are looking at the eﬀect of a drug on heart attacks, we might ﬁrst split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.15. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

Figure of blocking

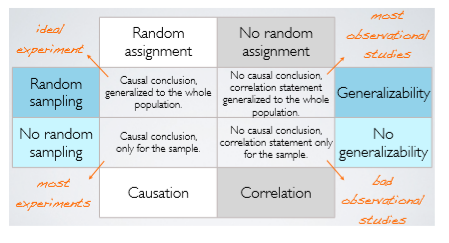




1.5.1. Difference between blocking and explanatory variables

* Factors are conditions we can impose on the experimental units.
* Blocking variables are characteristics that the experimental units come with, that we would like to control for.

**Random assignment vs. random sampling**



## 1.6 Examining numerical data

We examine numerical data by using:

* Scatterplots for paired data
* Dot plots and the mean
* Histograms and shape
* Variance and standard deviation
* Box plots, quartiles, and the median
* Robust statistics Transforming data
* Mapping data

### 1.6.1 Scatterplots

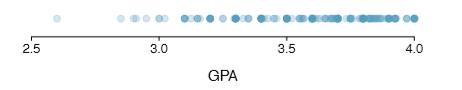
### **Scatterplots** are useful for visualizing the relationship between two numerical variables.

Do life expectancy and total fertility appear to be associated or independent?

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

### 1.6.2 Dot plots & mean

Useful for visualizing one numerical variable. Darker colours represent areas where there are more observations.



* The mean, also called the average, is one way to measure the centre of a distribution of data.
* The mean GPA is 3.59.

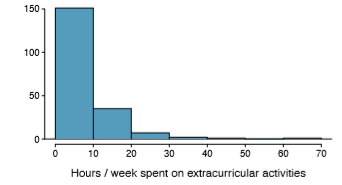
### 1.6.3 Histograms

* **Histograms** provide a view of the data density. Higher bars represent where the data are relatively more common.

Histograms are especially convenient for describing the shape of the data distribution.

We can take example of extracurricular Hours/ week spent on extracurricular activities.

* From histogram we can know the hours spent.



### 1.6.4 Variance and standard deviation

the variability in the data is so important. There are two measures of variability: **the variance and the standard deviation.**

Both of these are very useful in data analysis.

**The standard deviation** roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from it mean its deviation.

* **The variance** is roughly the average squared distance from the mean. **The standard deviation** is the square root of the variance. The standard deviation is useful when considering how close the data are to the mean.

**Their formulas**

* **The variance**:

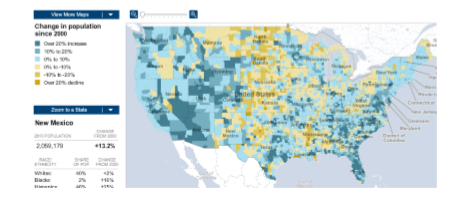


### 1.6.5 Intensity maps

* when we encounter geographic data, we should map it using an **intensity map**, where colours are used to show higher and lower values of a variable.

Figures 1.30 and 1.31 shows intensity maps for federal spending per capita (fed spend), poverty rate in percent (poverty), homeownership rate in percent (homeownership), and median household income (med income). The colour key indicates which colours correspond to which values.

**Fig: intensity map**



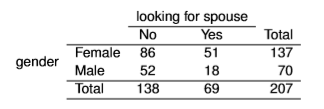
## 1.7 Considering categorical data

* Contingency tables and bar
* plots Row and column proportions
* Segmented bar and mosaic plots
* Pie charts
* Comparing numerical data across groups.

### 1.7.1 Contingency tables

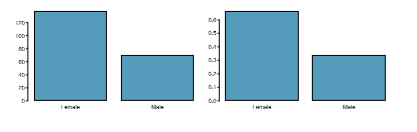
A table that summarizes data for two categorical variables is called a contingency table.

The contingency table below shows the distribution of students’ genders and whether or not they are looking for a spouse while in college.



### 1.7.2 Bar plots

* A bar plot is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a relative frequency bar plot.



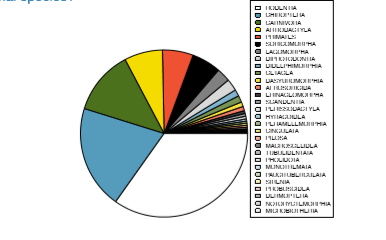
**How are bar plots different than histograms?**

* **Bar plots** are used for displaying distributions of categorical variables,

while **histograms** are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

### 1.7.3 Pie charts

* You can tell which order encompasses the lowest percentage of mammal species.



# **Chapter2: Probability**

## 2.1 probability

The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an inﬁnite number of times.

P(A) = Probability of event A

0 ≤ P(A) ≤ 1

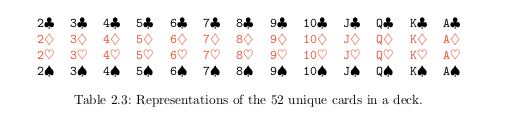
### 2.1.1 Disjoint or mutually exclusive outcomes

* Two outcomes are called disjoint or mutually exclusive if they cannot both happen.
* If A1 and A2 represent two disjoint outcomes, then the probability that one of them occurs is given by

**P (A1 or A2) = P(A1) + P(A2)**

### 2.1.2 Probabilities when events are not disjoint

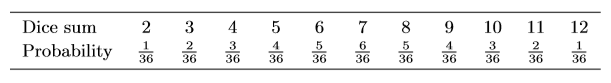
Let’s consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Table below:



### 2.1.3 Probability distributions

### A probability distribution is a table of all disjoint outcomes and their associated probabilities.

* Example: the probability distribution for the sum of two dice.



**Rules for probability distributions**

1. The outcomes listed must be disjoint.

2. Each probability must be between 0 and 1.

3. The probabilities must total 1.

### 2.1.4 Complement of an event

* Let us take an example of rolling a die the, **sample space (S)** = {1, 2, 3, 4, 5, 6}. Let **D = {2, 3}** represent the event that the outcome of a die roll is 2 or 3, **Then Complement of D (DC )** = {1, 4, 5, 6}.
* Simply **Complement of D is** the set of all possible outcomes not already included in D.

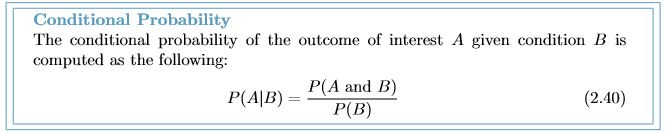
### 2.1.5 Independence

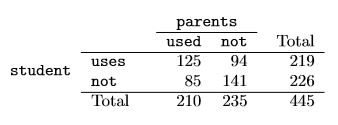
### Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

Example of rolling two dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die.

## 2.2 Conditional probability

## We call this a conditional probability because we computed the probability under a condition: parents = used. There are two parts to a conditional probability, the outcome of interest and the condition. It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event. We separate the text inside our probability notation into the outcome of interest and the condition:



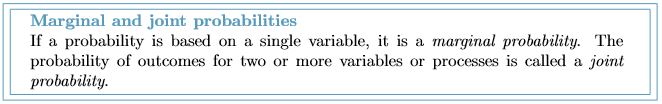


Example 2.35 A student is randomly selected from the study and she does not use drugs. What is the probability that at least one of her parents used? If the student does not use drugs, then she is one of the 226 students in the second row. Of these 226 students, 85 had at least one parent who used drugs:

P (parents = used given student = not) =85/226

= 0.376

### 2.2.1 Marginal and joint probabilities



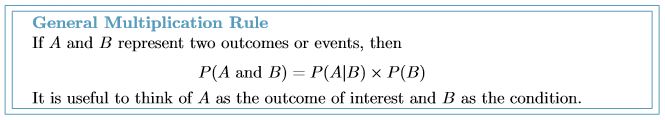
marginal probabilities for the sample, which are the probabilities based on a single variable without conditioning on any other variables. For instance, a probability based solely on the student variable is a marginal probability:

P student = uses) =219/445= 0.492

A probability of outcomes for two or more variables or processes is called a joint probability:

P (student = uses and parents = not) =94/ 445= 0.21

### 2.2.2 General multiplication rule



**Example 2.47:** Consider the smallpox data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

We will compute our answer using the General Multiplication Rule and then verify it using Table 2.16. We want to determine

P (result = lived and inoculated = no)

and we are given that

P (result = lived | inoculated = no) = 0.8588 P (inoculated = no) = 0.9608

Among the 96.08% of people who were not inoculated, 85.88% survived:

P (result = lived and inoculated = no) = 0.8588×0.9608 = 0.8251

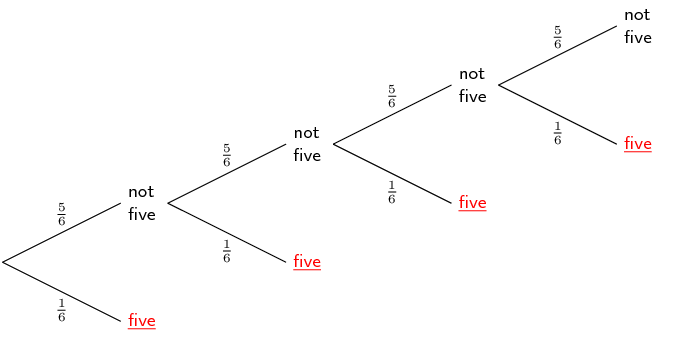
### 2.2.3 Independence considerations in conditional probability

If two processes are independent, then knowing the outcome of one should provide no information about the other. We can show this is mathematically true using conditional probabilities.

### 2.2.4 Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

**Example:**



### 2.2.5 Bayes’ Theorem

However, sometimes it is not possible to draw the scenario in a tree diagram. In these cases, we can apply a very useful and general formula: **Bayes’ Theorem.**

**Bayes’ Theorem:** inverting probabilities Consider the following conditional probability for variable 1 and variable 2: **P(outcome A1 of variable 1 | outcome B of variable 2)**

Bayes’ Theorem states that**: ”this conditional probability can be identiﬁed as the following fraction: P(B|A1)P(A1)/ P(B|A1)P(A1) + P(B|A2)P(A2) +···+ P(B|Ak)P(Ak)”**

where A2, A3, ..., and Ak represent all other possible outcomes of the ﬁrst variable.

**Note:** Only use **Bayes’ Theorem** when tree diagrams are diﬃcult Drawing a tree diagram makes it easier to understand how two variables are connected. Use Bayes’ Theorem only when there are so many scenarios that drawing a tree diagram would be complex.

**Example 2.61** Professors sometimes select a student at random to answer a question. If each student has an equal chance of being selected and there are 15 people in your class, what is the chance that she will pick you for the next question?

If there are 15 people to ask and none are skipping class, then the probability is 1/15, or about 0.067.

**Example 2.62** If the professor asks 3 questions, what is the probability that you will not be selected? Assume that she will not pick the same person twice in a given lecture.

For the ﬁrst question, she will pick someone else with probability 14/15. When she asks the second question, she only has 14 people who have not yet been asked. Thus, if you were not picked on the ﬁrst question, the probability you are again not picked is 13/14. Similarly, the probability you are again not picked on the third question is 12/13, and the probability of not being picked for any of the three questions is

P(not picked in 3 questions) = P(Q1 = not picked, Q2 = not picked, Q3 = not picked.)

=14/15 ×13/14 ×12/13

=12/15= 0.80

## 2.3 Random variables

Random variable is a random process or variable with a numerical outcome.

We call a variable or process with a numerical outcome a random variable, and we usually represent this random variable with a capital letter such as X, Y , or Z. The amount of money a single student will spend on her statistics books is a random variable, and we represent it by X.

**There are two types of random variables:**

**Discrete random variables** often take only integer values

Example: Number of credit hours, Difference in number of credit

hours this term vs last

**Continuous random variables** take real (decimal) values

Example: Cost of books this term, Difference in cost of books this

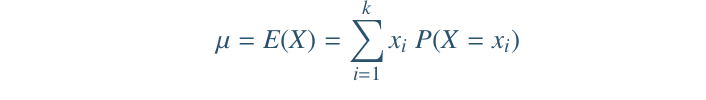
term vs last

We are often interested in the average outcome of a random

variable.

We call this the expected value (mean), and it is a weighted

average of the possible outcomes



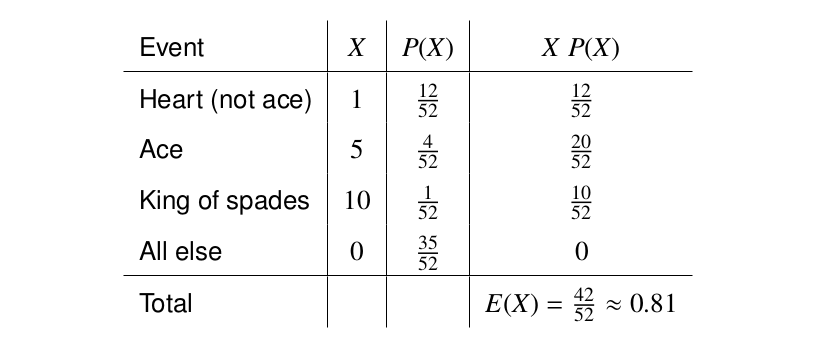
**Example:**

In a game of cards you win $1 if you draw a heart, $5 if you draw an

ace (including the ace of hearts), $10 if you draw the king of spades

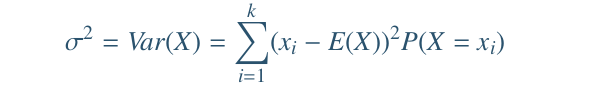
and nothing for any other card you draw. Write the probability model

for your winnings, and calculate your expected winning.

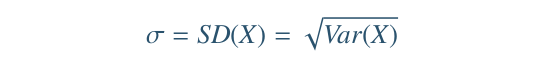


### 2.3.1 Variability in random variables

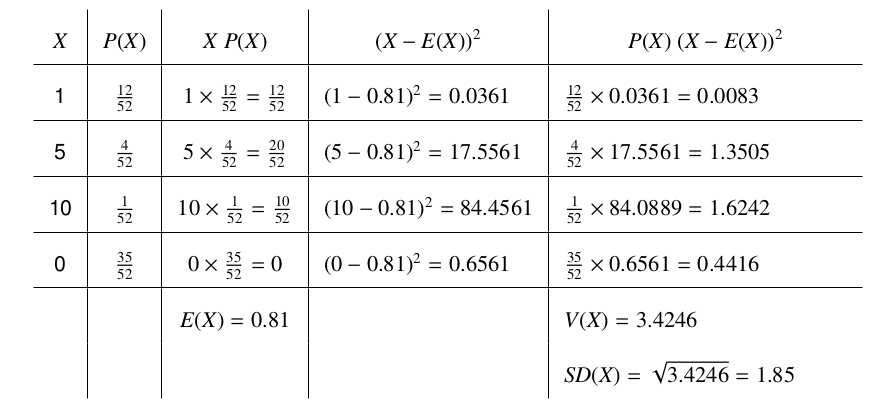
* The variance and standard deviation can be used to describe the variability of a random variable.
* General variance formula If X takes outcomes x1, ..., xk with probabilities P(X = x1), ..., P(X = xk) and expected value µ = E(X), then the variance of X, denoted by V ar(X) or the symbol σ2,



* The standard deviation of X, labeled σ, is the square root of the variance.



**Example:**



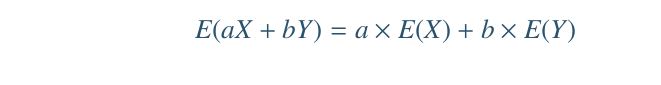
### 2.3.2Linear combinations of random variables

* So far, we have thought of each variable as being a complete story in and of itself. Sometimes it is more appropriate to use a combination of variables. For instance, the amount of time a person spends commuting to work each week can be broken down into several daily commutes. Similarly, the total gain or loss in a stock portfolio is the sum of the gains and losses in its components.
* A linear combination of two random variables X and Y is a fancy phrase to describe a combination

aX + bY

* Where a and b are some fixed and known numbers
* The average value of a linear combination of random variables is

given by:



**Example:**

On average you take 10 minutes for each statistics homework problem

and 15 minutes for each chemistry homework problem. This week

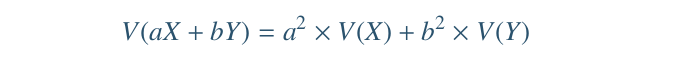
you have 5 statistics and 4 chemistry homework problems assigned.

What is the total time you expect to spend on statistics and physics

homework for the week?

### 2.3.3 Variability in linear combinations of random variables

* Quantifying the average outcome from a linear combination of random variables is helpful, but it is also important to have some sense of the uncertainty associated with the total outcome of that combination of random variables
* variance of a linear combination of random variables can be computed by plugging in the variances of the individual random variables and squaring the coeﬃcients of the random variables:

The standard deviation of the linear combination may be found by taking the square root of the variance.

**Example:**

The standard deviation of the time you take for each statistics home-

work problem is 1.5 minutes, and it is 2 minutes for each chemistry

problem. What is the standard deviation of the time you expect to

spend on statistics and physics homework for the week if you have 5

statistics and 4 chemistry homework problems assigned?

### 

**END !!!!!**