Essential Mathematics for BioNLP

Harsh Thakkar and Prasenjit Majumder

Harsh Thakkar
DA-IICT, Gandhinagar, Gujarat, India - 382007 e-mail: harsh9t@gmail.com

Prasenjit Majumder
DA-IICT, Gandhinagar, Gujarat, India - 382007 e-mail: Prasenjit.mamjumder@gmail.com

Contents

Abstract Each chapter should be preceded by an abstract (10–15 lines long) that summarizes the content. The abstract will appear *online* at www.SpringerLink.com and be available with unrestricted access. This allows unregistered users to read the abstract as a teaser for the complete chapter. As a general rule the abstracts will not appear in the printed version of your book unless it is the style of your particular book or that of the series to which your book belongs.

Please use the 'starred' version of the new Springer `abstract` command for typesetting the text of the online abstracts (cf. source file of this chapter template `abstract`) and include them with the source files of your manuscript. Use the plain `abstract` command if the abstract is also to appear in the printed version of the book.

1 Introduction: *Mathematics is everywhere!*

Yes, one may believe it or not, but the universal truth is that there are numbers and they are everywhere. The reality we define is the product of our own perception and the outcome of the actions we take. In mathematical terms everything happening around can be defined as functions and the relations between them. Such a concept also applies to the biology domain. In this chapter we focus on building up the basic mathematical concepts as a necessity to thoroughly understand the analytical calculation based challenges in the biology domain. This chapter aims to bring the biologists to a common ground in terms of mathematical understanding of problems as compared to computer science professionals. We cover the essential blocks of mathematical and statistical analysis involved in biological domain by a brief study of topics like basic set theory, probability, and statistics.

2 Basics of Set Theory

2.1 Introduction

What is a set?
Set: A *set* in general is a collection of objects or items. Formal definitions of a set are very scarce as it is considered to be a very trivial concept. For instance, a set can be *the set characters in the English alphabet* or *a collection of the all the species of mammals on earth* and also *the group of all the female students in the class*. Sets may consist of numbers, alphabetic characters, and even equations. Thus, sets are of very precise or definite in nature. One can also conclude from the above examples that a set can be formed by the objects, items or elements which have something common in nature.

2.1.1 Notations:

Mathematically, sets are often represented as a English capital alphabets or characters e.g. **A, B, S, X ,Y, Z,** and etc., while its elements are represented by small English alphabets or characters e..g **a,b,c,x,y,z,** and etc. The mathematical notation for a set is as:

$$S = \{a, b, c, d\} \tag{1}$$

To make it clear let us consider an example for biology, viz. the fundamental elements of DNA. Since the DNA helix is a pattern formed by an alternating sequence of either of Adenine (A), Thymine (T), Cytosine (C), or Guanine (G) nucleobases. The set D consists of the elements A,T,C,G, which is mathematically represented as:

$$D = \{a, t, c, g\}. \tag{2}$$

There exist a variety of formats which are universally accepted for representing sets. Here we list a few of the standardized notations used for representing sets.

1. **List notation**: This notation is the same as we saw earlier in equation 2. Here this set is represent by the list of elements within curly braces.
2. **Predicate notation**: In this notation a set is represented in terms of the qualifying condition or property that the members of the set have in common.
   $\{x \mid P(x)$; P is some predicate (i.e. a condition or property that has to be satisfied by the members of the set)$\}$ Here the $\mid$ operator is read as *such that*. For instance, $\{x \mid P(x)$; x is a odd integer and x $\geq 8\}$. This would be the set say $S = \{1,3,5,7\}$.

3. **Recursive/Rules notation**: This is a very primitive style of representing sets, In this style a pseudo algorithm format is used for presenting the set e.g. Consider a set A, such that all elements of set A are ≥ 3. This set is represented as:

   a. $4 \in E$
   b. if x $\in E$, then $x+2 \in E$
   c. nothing else $\in E$

   Thus, there exists a variety of styles for representing sets accepted by practitioners. The predicate style is commonly used by mathematicians and professors. While the others are used by other interdisciplinary personnel.

2.1.2 Operations on sets:

[from settheory.pdf] A variety of operations are defined on the sets. Consider two sets A and B with elements a and b respectively. The operation that are defined on these sets are:

- **Basic operations:** The basic operations on sets are-
  - $a \in A$: There exists elements that are the subsets of the defining set A.
  - $a \notin A$: There exists elements that are NOT the subsets of the defining set A.
  - $A \subset B$: There may exist a relation between two sets say, A and B such that A is a subset of B. e.g. Let set A = $\{$The set of all odd integers$\}$ and set B = $\{$The set of all integers$\}$. Thus A $\subset$ B.
  - $A \cap B$: There may exist some common elements between two sets A and B.
  - $A \cup B$: The exists a relation such that two elements of sets can be merged together to form a new set. The Union operation can also be perceived as *addition* in parallel mathematical notation.
  - $A \cap B = \{\phi\}$: There may exists sets which have no element in common or the element common to both the sets is the *null* element or also known as *zero* element.
- **Additional operators:**

- $\bar{A}$ or A^c or $\grave{A}$: This is called the compliment operator. In order to define a complement of a set, we have to first understand the notion of a universal set. All sets are subsets of a universal set, generally presented by symbol $\bigcup$. A compliment of a set A is the set of all the elements except those present in set A. e.g. Lets define universal set $\bigcup$ = {fruits, flowers, insets, mammals, reptiles} and let there be a set A = {fruits, insects}. Then, $\bar{A}$ will be the set of all items except fruits and insects viz. $\bar{A} = \{flowers, mammals, reptiles\}$. It can also be represented mathematically as $\bar{A}$ = $\bigcup$ - A.
- $A - B$: This $-$ is the called the difference operator. It calculates the difference between two sets. Mathematically, it is defined as x| x $\in$ A and x $\notin$ B; i.e. the set all elements that are unique to only set A. With respect to the above complement example, lets define a set B = {mammals, fruits, reptiles} Here, $A - B$ = {insects}.

- $P(A)$: Power Set, is a set which consists of all the subsets of a set A including the null set $\{\phi\}$ and the set A itself. For instance, let set A = {set of all integers $\geq$ 3} then, power set of A, P(A) = {{ϕ}, {1}, {2}, {3}, {1,2}, {1,3}, {2,3}, {A}}.

2.1.3 Fundamental laws of set theory

The set theory is founded on some fundamental laws such as:

1. **Idempotent law**:
   $A \cup A = A$;
2. **commutative law**:
   $A \cup B = B \cup A$ and $A \bigcap B = B \bigcap A$
3. **Associative law**:
   $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$
4. **Distributive law**:
   $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \bigcap (B \cup C) = (A \cap B) \cup (A \bigcap C)$
5. **Complement law**:
   $A \cup A = \bigcup$ (universal set);
   $\bar{\bar{A}} = A$;
   $A \cap \bar{A} = \{\phi\}$;
   $A - B = A \cap \bar{B}$
6. **Identity law**:
   $A \cup \phi = A$
   $A \cap \phi = A$
   $A \cup \bigcup = \bigcup$
   $A \cap \bigcup = A$
7. **De Morgan's law**:
   $(\bar{(A \cap B)}) = \bar{A} \cup \bar{B}$ and $(\bar{A \cup B}) = \bar{A} \cap \bar{B}$
8. **Consistency principle**:
   $A \subset B$ if and only if $A \cup B = B$ **and** $A \cap B = A$

3 Basics of Probability

Probability, as the name suggests is the possibility of something happening which is expected. It is also known as the likelihood of a specific result in a experiment. In real life, we often use the word *probability* related to the degree of expectancy of circumstances to occur. In mathematics, situations are conditions and the outcomes are events. The occurring of events is random and probability is the measure of randomness of events.

According to the classical definition of probability:
The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.[1]

We often quantify the randomness of an event through constructing a mathematical model and then calculate the odds of a precedented outcome. The probability theory is inspired by real life decisions. We either do something or we do not do it, the decisions can be described in binary i.e. 1=yes and 0=no. However, not always we are certainly sure about what we want to do. There are a various factors that cloud our judgment or have an amount of impact on it. Similarly, this vagueness is reflected in mathematical probability theory as it varies from **0 to 1**. The value of an event happening varies from $\{0.0,..,0.01,..,0.1,..,1\}$.

Mathematically:
Consider a un-biased coin. Let P be the set2.1 of all possible outcomes of tossing that coin once. Thus, the out could be either a *head*[H] or a *tail*[T].
$P = \{H,T\}$ Here, the total number of possible events is 2. Hence, the probability of getting a head an outcome will be:

$$Probablility = \frac{\text{Total number of heads}}{\text{Total number of outcomes}} = \frac{1}{2} \tag{3}$$

Similarly the probability of getting a tail is also $\frac{1}{2}$.

Now let us consider there are two such u-biased coins simultaneously, and we want to find out the probability of *at-least one head occurring*. Here, the set of total possible events, say P, will be:
$P = \{\{H,H\},\{H,T\},\{T,H\},\{T,T\}\}$
Now, probability of getting at-least one head in the outcome will be equal to the number of elements in set P which have at-least one or more H. Thus,

$$Probablility = \frac{\text{Outcomes with one one more heads}}{\text{Total number of outcomes}} = \frac{3}{4} \tag{4}$$

whereas, the probability for getting exactly two heads, or both heads will be $\frac{1}{4}$.

3.1 Fundamental laws of probability theory

There are certain fundamental laws, properties or rules that are followed in probability theory.

- *The total probability of all possible outcomes of an event will be always equal to 1*. Since, the total probability of all outcomes is their sum, the total value of the set is the set itself, say P. One cannot have a probability value say 10 or 100 or 540. An event either happens or does not. Just as a person cannot be alive and dead at the same time, bound by the physical laws of nature!
- $P(\phi) = 0$ - Probability of no event happening is nothing i.e. 0.
- $P(\bar{A}) = 1 - P(A)$ - The probability of an event not happening is 1 minus the probability of that event happening. This is also called the complement probability as studies in earlier set theory (section 2.1.3).
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ - The probability of two events happening together is calculated by this equation. It can be understood better by the following diagram:

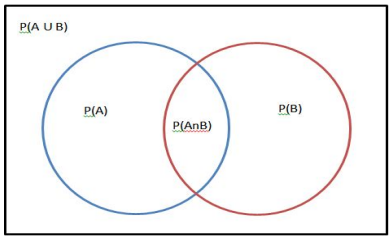

Fig. 1 The Venn diagram displaying the probability of two events A and B.

- *if* $A \subset B$ *then* $P(A) \leq P(B)$ - for instance, consider the example from section 2.1.2. Here, set A = {fruits, insects} and set $\bigcup$ = {fruits, flowers, insets, mammals, reptiles}. Clearly, $A \subset \bigcup$ as $\bigcup$ is the universal set it has more number of elements than that of set A. Thus $P(A) \leq P(\bigcup)$.
- *Independent events* - For these events, the outcome of one event doesn't have an effect or influence over the outcome of other event. For instance, recall the earlier problem of flipping two coins. Here the probability of getting two heads was $\frac{1}{4}$, since the outcome of one coin is independent of the outcome of the other coin. Thus, for independent events $P(A \cap B) = P(A) * P(B)$
- *Mutually exclusive events* - Two events are said to me mutually exclusive if they both cannot occur together. Either of them can occur but not both. For instance, a professor p can be present either in the auditorium or in the office at a given point of time, but not both. Also, the probability of getting a head or a tail on single toss of an un-biased coin is $\frac{1}{2}$, it is not possible to have a head and a tail at the same time. The vertical standing coin case is not taken into account!

3.1.1 Random Variables

3.2 Conditional Probability and Bayes theorem

Conditional probability is the probability of an event occurring based on the outcome/condition of occurring of some other event. For instance, say P(X) be the probability a injection syringe being re-used and P(Y) be the probability of catching HIV. Here, both the probabilities are related. P(Y)'s value depends on the value of P(X). Mathematically, the notation used for conditional probability is $P(E|H)$, i.e. probability of event Y occurring given that event X occurs.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{5}$$

when, $P(B) \geq 0$

$$P(X|Y) = 0 \tag{6}$$

when, $P(B) = 0$

$$P(X|Y) = P(X) \tag{7}$$

To understand this concept lets consider an example of two buckets, B1 with 10, 60, and 30 red, white, and blacks balls, B2 with 10, 40, and 50 red, white, and black balls respectively. We choose one ball randomly out of the two buckets. Consider the following sample problems.

1. **What is the probability that bucket B2 is selected and a white ball is chosen?** i.e. $P(B2|W)$
   Here,
   $$P(B2|W) = \frac{P(W)*P(B2)}{P(B2)} = \frac{40/100}{1/2} = \mathbf{0.2}$$
2. **What is the probability of choosing a red ball and it belongs to bucket B1?** i.e. $P(R|B1)$?
   Here,
   $$P(B1|R) = \frac{P(R \cap B1)}{P(B1)} \text{ (from equation 5)}$$
   lets first calculate the individual probabilities:
   $$P(R,B1) = \frac{P(R \cap B1)}{P(B1)} = \frac{P(R)*P(B1)}{P(B1)} = \frac{5/100}{1/2} = \mathbf{0.1}$$
   Similarly, the other probabilities will be:
   $$P(W|B1) = 0.3 \; P(B|B1) = 0.15 \; P(R|B2) = 0.05$$
   $$P(W|B2) = 0.2 \; P(B|B2) = 0.25$$
   Thus,
   $P(R|B1) = \frac{P(R \cap B1)}{P(B1)} = \frac{10/200}{1/2} = \frac{0.05}{0.5} = \mathbf{0.1}$

3. **What is the probability that *either* bucket B1 is chosen *or* the ball is red?** i.e. $P(B1 \cup R)$?
   We know that,

$$P(B1 \cup R) = P(B1) + P(R) - P(B1 \cap R)$$

   Therefore,

$$P(B1 \cup R) = P(B1) + P(R) - [P(B1) * P(R)]$$
$$= \frac{1}{2} + \frac{20}{200} - [(\frac{1}{2}) * (\frac{20}{200})] = 0.5 + 0.1 - 0.05 = \mathbf{0.55}$$

Thus, in this section we learned about conditional probability and its application with some thorough sample practice examples. In the next section we will learn about the very famous bayes theorem, which the outcome of exhaustive study of conditional probability theory, in brief.

3.2.1 Bayes theorem

In real life, the outcomes of events change with the change in their initial situations. So is in the mathematics, in the probability. Probability of any event is based on the evidences or occurrence of its previous events. The flow of a stream will vary according to the steepness of the hill or mountain. There is a relation between the slope of the mountain and the speed that flowing stream acquires during its decent to the sea. Such relations need mathematical explanations and hence, bayes theorem came into existence.

The **Bayes theorem** is called so, after the name of famous statistician and philosopher Reverned Thomas Bayes. This theorem has been quoted as:
`Bayes is to probability as Pythagoras is to geometry` by Sir Harrold Jefferys[2]. The main purpose of the bayes theorem is to analyze the relations between evidences and their theories. Bayes theorem is an important finding as it helps us understand precisely the statistical inference and predict the events based on prior evidences.

Bayes Rule, which suggests that the probability of a hypothesis given a certain evidence, i.e. the posterior probability of a hypothesis, can be obtained in terms of the prior probability of the evidence, the prior probability of the hypothesis and th conditional probability of the evidence given the hypothesis. Mathematically,

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(E)} \tag{8}$$

where,
$P(H|E)$- posterior probability of the hypothesis.
$P(H)$- prior probability of hypothesis.
$P(E)$- prior probability of evidence.
$P(E|H)$- conditional probability of evidence of given hypothesis.
Or in a simpler form:

$$Posterior = \frac{(Prior) \times (Likelihood)}{Evidence} \tag{9}$$

Lets, see a couple of practice examples to understand this concept. Consider the following examples:

- **The HIV Test**: A study is undertaken comparing the effectiveness of a HIV test on a population of patients. The information regarding the HIV status of the population is known apriori. Lets say part of population which is HIV positive is X and HIV negative is $\bar{X}$. It is provided that the HIV tests have 99.7% of correctly identifying HIV positive patients, and 98.5% of correctly identifying HIV negative patients. Assume that a patient from the population with a 0.1% prevalence of HIV, receives a positive test result. What is the probability that this patient is actually infected with HIV? i.e. $P(X|+)$?
  Here, Re-writing the given information from the question:
  $P(+|X) = 0.997$, $P(-|\bar{X}) = 0.985$, P(X) = 0.01, $\bar{X} = 1 - X$
  From equation 3.2.1, we can write:

$$P(X|+) = \frac{P(+|X)P(X)}{P(+|X)P(X) + P(+|\bar{X})P(\bar{X})} \tag{10}$$

$$= \frac{P(+|X)P(X)}{P(+|X)P(X) + [1 - P(-|\bar{X})][1 - P(\bar{X})]} \quad \text{(given)}$$

$$= \frac{0.997 * 0.001}{0.997 * 0.001 + 0.015 * 0.999}$$

$$\mathbf{= 0.062}$$

  Thus, given the population, in this test the patient has only a 6% probability of being infected by HIV disease.

- **Predicting membrane proteins** In their work [3], Peter woolf et.al., have presented a problem regarding classifying the membrane proteins using the fraction of hydrophobic residues. The problem is stated as:
  A researcher hypothesizes that it is possible to detect membrane proteins using the fraction of hydrophobic residues alone. To test this model, the researcher creates a library of 7500 proteins and scores each of these proteins based on their fraction of hydrophobic residues and whether they are membrane proteins. Given the results of this analysis below, what is the probability that a novel protein that is primarily hydrophobic is also a membrane protein? i.e. $P(B|A)$?

	Majority hydrophobic	Majority hydrophilic
Membrane bound	2911	961
Cytosolic	713	2915

  Lets first calculate the basic probabilities based on the given data from the above table. We summarize the data as:
  Majority hydrophobic = A, Majority hydrophilic = $\bar{B}$

Membrane bound = B, Cytosolic = $\bar{B}$

	Majority hydrophobic	Majority hydrophilic
Membrane bound	$P(A\|B)$	$P(\bar{A}\|B)$
Cytosolic	$P(A\|\bar{B})$	$P(\bar{A}\|\bar{B})$

Thus, from given table we can calculate the basic probabilities, for instance:

$P(A|B) = \frac{2911}{7500} = 0.388$

Similarly, $P(\bar{A}|B) = 0.128$, $P(A|\bar{B}) = 0.095$, $P(\bar{A}|\bar{B}) = 0.389$

Thus, $P(B|A) = \frac{P(A|B)}{P(A)} = \frac{P(A\cap B)/P(B)}{P(A)}$
$= \frac{0.388}{0.483} =$ **0.803**

Hence, the probability of a novel protein being membrane bound given that it is hydrophobic is 0.803 or 80.3%.

- **The Raven's Paradox** Also very famous as Hempel's paradox after the name of its proposer Carl Gustav Hempel, questions the very intuition on the concept of what to consider as an evidence for an event. It became famous in the 1940's as an illustration of contradiction between the theory of inductive logic and intuition. This problem is rather more philosophical than numerical. It is included to demonstrate the wide range of application of the bayes theorem.
  The paradox is proposed in terms of hypothesis [4, 5, 6, 7]:

  1. `All ravens are black.`
  2. On the basis of strict logic. the above statement is equivalent to
     `Everything that is not black is not a raven.`
  3. From the above statements it can be understood that whenever second statement is true, so is the first one. Similarly, always when second statement is false, so is the first one (imagining a world where something that isn't black and yet is a raven exists). Thus, any statement like:
     `Nevermore, my pet raven, is black.` supports the hypothesis of the first statement that all ravens are black.
  4. The contradiction occurs when the above process is applied to a statement like *This red (and hence not black) thing is a cherry (and thus not a raven.)* on observing a red cherry. By the same logic this statement becomes an evidence to the above statement that `All ravens are black` and also supports `Everything that is not black is not a raven`. This information cannot be considered subtle as it appears that the information on ravens has been obtained, by the theory of induction, observing cherries.
     The exist many solutions to avoid the paradox ,i.e. violation of intuition, from being formulated. One of them is by using the bayes theorem. Whilst using bayes theorem the above paradox wouldn't have arrived instead. Since, evi-

dence (i.e. the hypothesis), in bayes rules, is calculated/tested by multiplying a the ratio:

$$\frac{\text{probability of observing that A is true given B}}{\text{probability of observing A}}$$

Hence, when selecting a cherry at random and observing it the probability of observing a red cherry is independent of the color of ravens. The instance when numerator will be equal to demonstrator, the value will become one and the probability will remain unaltered. Seeing a red cherry won't affect your belief about whether all ravens are black or not. If you ask someone to select a non-black-thing at random, and they show you a red cherry, then the numerator will exceed the denominator by a negligible amount. Observing the red cherry will only slightly increase your evidence (i.e. the hypothesis) that all ravens are black. You'll have to see almost every non-black-thing in the world (and see they're all non-ravens) before your hypothesis statement one that `All ravens are black` increases substantially. This result in both cases, supports and satisfies the intuition.

- **The Blind person**: The ratio of color blind men in a country is 2:100, and for women it is 1:1000. The population consists of 47% of women. If a person is selected at random from the total population, then;

  1. What is the probability that a person chosen at random is color blind?
  2. What is the probability that a person chosen at random is a man given that he is color blind?

  Here, based on the provided information let us first summarize the data as:

  Ratio of men in population = $\frac{53}{100}$,
  Ratio of women in population = $\frac{47}{100}$,
  Ratio of total color blind men = (ratio of men)*(ratio of color blind men) = $\frac{53}{100} * \frac{2}{100} = 0.101107$,
  Ratio of total color blind women = (ratio of women)*(ratio of color blind women) = $\frac{47}{100} * \frac{0.1}{100} = 0.00047$

  1. The probability that a person chosen at random is color blind is:
     $P(colorblind)$ = (ratio of total color blind men + ratio of total color blind women) = 0.101107 + 0.00047 = **0.101107**
  2. the probability that a person chosen at random is a man given that he is color blind is:

     $P(man|colorblind) = \frac{P(colorblind|man)*P(man)}{P(colorblind)}$

     $= \frac{(0.53)(0.02)}{0.01107}$ = **0.9575**

For further study of examples related to biology please refer the work [9], where the authors William and Matthew present the use of bayesian statistics for pedigree analysis, yet another application of bayes theorem.

4 Basics of Statistics

What is statistics?
We define this term classifying it into a set of definitions based on the perceptions of various genre of people using it.

- Generally: Statistics is a field of mathematics that pertains to data analysis. Statistical methods and equations can be applied to a data set in order to analyze and interpret results, explain variations in the data, or predict future data. A few examples of statistical information we can calculate are mean (average), median (mid-value), mode (most repeated value).
- In the mind of a statistician, the world consists of populations and samples. An example of a population is all 7th graders in the United States. A related example of a sample would be a group of 7th graders in the United States. In this particular example, a federal health care administrator would like to know the average weight of 7th graders and how that compares to other countries. Unfortunately, it is too expensive to measure the weight of every 7th grader in the United States. Instead statistical methodologies can be used to estimate the average weight of 7th graders in the United States by measure the weights of a sample (or multiple samples) of 7th graders.
- In a simplified manner: Statistics a set of concepts, rules, and procedures that help us to:
  - Organize numerical information in the form of tables, graphs, and charts.
  - Understand statistical techniques underlying decisions that affect our lives and well-being.
  - take educated decisions.

In the upcoming sections we will discuss and examine a variety of measures used in statistics pertaining to biology and in general domain. But before getting in to the deep waters of dizzy ocean, let us first understand some concepts and terminologies used in statistics.

- **Data**: Data can be summarized as facts, observations, and information that come from experiments and investigations.

  Measurement data sometimes called quantitative data – the result of using some instrument to measure something (e.g., test score, weight);

  Categorical data also referred to as frequency or qualitative data. Things are grouped according to some common property/ies and the number of members of

the group are recorded (e.g., males/females, vehicle type).

- **Variables**: A variable can be described as a property of an object or event that can take on different values. For example, college major is a variable that takes on values like mathematics, computer science, English, psychology, etc.

  Discrete Variable - a variable with a limited number of values (e.g., gender (male/female), college class (freshman/sophomore/junior/senior). Various categories of variables are:

  Continuous Variable - a variable that can take on many different values, in theory, any value between the lowest and highest points on the measurement scale.

  Independent Variable - a variable that is manipulated, measured, or selected by the researcher as an antecedent condition to an observed behavior. In a hypothesized cause-and-effect relationship, the independent variable is the cause and the dependent variable is the outcome or effect.

  Dependent Variable - a variable that is not under the experimenter's control – the data. It is the variable that is observed and measured in response to the independent variable.

  Qualitative Variable - a variable based on categorical data.

  Quantitative Variable - a variable based on quantitative data.

- **Graphs**: These are the visual display of data used to present frequency distributions so that the shape of the distribution can easily be seen. There exist a variety of graphs, the most frequently used types of graphs are:
  Bar graph - a form of graph that uses bars separated by an arbitrary amount of space to represent how often elements within a category occur. The higher the bar, the higher the frequency of occurrence. The underlying measurement scale is discrete (nominal or ordinal-scale data), not continuous.

  Histogram - a form of a bar graph used with interval or ratio-scaled data. Unlike the bar graph, bars in a histogram touch with the width of the bars defined by the upper and lower limits of the interval. The measurement scale is continuous, so the lower limit of any one interval is also the upper limit of the previous interval.

  Box-plot - a graphical representation of dispersions and extreme scores. Represented in this graphic are minimum, maximum, and quartile scores in the form of a box with "whiskers." The box includes the range of scores falling into the middle 50% of the distribution (Inter Quartile Range = 75th percentile - 25th percentile)and the whiskers are lines extended to the minimum and maximum scores in the distribution or to mathematically defined $(\pm 1.5 * IQ_R)$ upper and

lower fences.

Scatter-plot - a form of graph that presents information from a bivariate distribution. In a scatterplot, each subject in an experimental study is represented by a single point in two-dimensional space. The underlying scale of measurement for both variables is continuous (measurement data). This is one of the most useful techniques for gaining insight into the relationship between tw variables.

Statistical measures related to graphs:

– Range: The simplest measure of variability to compute and understand is the range. The range is the difference between the highest and lowest score in a distribution. Although it is easy to compute, it is not often used as the sole measure of variability due to its instability. Because it is based solely on the most extreme scores in the distribution and does not fully reflect the pattern of variation within a distribution, the range is a very limited measure of variability.

– Quartile: In descriptive statistics, the quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. A quartile is a type of quantile. The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set. The second quartile (Q2) is the median of the data. The third quartile (Q3) is the middle value between the median and the highest value of the data set. In applications of statistics such as epidemiology, sociology and finance, the quartiles of a ranked set of data values are the four subsets whose boundaries are the three quartile points. Thus an individual item might be described as being ”in the upper quartile”.

  1st quartile (designated Q1) also called the lower quartile or the 25th percentile (splits off the lowest 25% of data from the highest 75%)
  2nd quartile (designated Q2) also called the median or the 50th percentile (cuts data set in half)
  3rd quartile (designated Q3) also called the upper quartile or the 75th percentile (splits off the highest 25% of data from the lowest 75%)
  Interquartile Range (designated IQ_R) is the difference between the upper and lower quartiles. (IQR = Q4 - Q2)

  For instance, considered the sample ordered data: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49
  Total number of item in this data set is = 11. Hence, Q1 = 15, Q2=40, Q3 = 43

4.1 Mean, Median and Mode

The mean, median and mode are known as the measures of center[13], since the purpose of their use is to find the center of the distribution of values. To understand these three measure lets consider an example illustrated by McDonal in his work [10], the problem is stated as-

The Maryland Biological Stream Survey used electro-fishing to count the number of individuals of each fish species in randomly selected 75-m long segments of streams in Maryland. Here are the numbers of black-nose dace, Rhinichthys atratulus, in streams of the Rock Creek watershed:

Mill_Creek_1	76
Mill_Creek_2	102
North_Branch_Rock_Creek_1	12
North_Branch_Rock_Creek_2	39
Rock_Creek_1	55
Rock_Creek_2	93
Rock_Creek_3	98
Rock_Creek_4	53
Turkey_Branch	102

- **Mean**: The mean is the most common measure of central tendency and the one that can be mathematically manipulated. It is defined as the average of a distribution. Mean is represented by a bar over the variable i.e. $\bar{X}$.

$$\bar{X} = \frac{\sum X_i}{N} \tag{11}$$

  Thus, the mean is nothing but the sum of all the scores in the distribution (X_i) divided by the total number of scores (N). The mean is the balance point in a distribution such that if you subtract each value in the distribution from the mean and sum all of these deviation scores, the result will be zero. In the above example the mean will be equal to:

$$\bar{X} = \frac{\sum X_i}{N} = \frac{76+102+12+39+55+93+98+53+102}{9} = 70.0$$

- **Median**: The median is the score that divides the distribution into halves; half of the scores are above the median and half are below it when the data are arranged in numerical order. The median is also referred to as the score at the 50th percentile in the distribution. The median location of N numbers can be found by the formula $\frac{(N+1)}{2}$. When N is an odd number, the formula yields a integer that represents the value in a numerically ordered distribution corresponding to the median location. (For example, in the distribution of numbers (3 1 5 4 9 9 8) the median location is $\frac{(7+1)}{2} = 4$. Thus, the 4^{th} value is the median, i.e. 4. When applied to the ordered distribution (1 3 4 5 8 9 9), the value 5 is the median, three

scores are above 5 and three are below 5. If there were only 6 values (1 3 4 5 8 9), the median location is $\frac{(N/2)+[(N/2)+1]}{2}$ = 3.5. Here, the median is half-way between the 3rd and 4th scores (4 and 5), i.e. 4.5. Thus, in our earlier example:

Number of data values = 9
Sorted data = 12,39,53,55,76,93,98,102,102
Median = $\frac{(9+1)}{2}$ = 5
The 5^{th} item is the median, i.e. = 76

- **Mode**: The mode of a distribution is simply defined as the most frequent or common score in the distribution. If the highest frequency is shared by more than one value, the distribution is said to be multimodal. It is not uncommon to see distributions that are bimodal reflecting peaks in scoring at two different points in the distribution. In the example discussed before, there is only one value that is frequent or repeats itself the most, i.e. 102.

  Thus, mode = 102

4.2 Variance and Standard deviation

These measures are also known as the measures of spread. Although the average value in a distribution is informative about how scores are centered in the distribution, the mean, median, and mode lack context for interpreting those statistics. Measures of variability provide information about the degree to which individual scores are clustered about or deviate from the average value in a distribution.

- **Variance**: The variance is a measure based on the deviations of individual scores from the mean. As noted in the definition of the mean, however, simply summing the deviations will result in a value of 0. To get around this problem the variance is based on squared deviations of scores about the mean. A variance of zero indicates that all the values are identical. Variance is always non-negative: a small variance indicates that the data tends to be very close to the mean (expected value) and hence to each other, while a high variance indicates that the data is very spread out around the mean and from each other. The sum of the squared deviations, $\sum(X_i - \bar{X})$, is divided by N (population) or by N - 1 (sample). The result is the average of the sum of the squared deviations and it is called the variance.

- **Standard deviation**: An equivalent measure is the square root of the variance, called the standard deviation. The standard deviation has the same dimension as the data, and hence is comparable with deviations of the mean. The standard deviation (denoted by σ) is defined as the positive square root of the variance.

The variance is a measure in squared units and has little meaning with respect to the data. Thus, the standard deviation is a measure of variability expressed in the same units as the data. The standard deviation is very much like a mean or an *average* of these deviations. In a normal (symmetric and mound-shaped) distribution, about two-thirds of the scores fall between $[+1, -1]$ standard deviations from the mean and the standard deviation is approximately 1/4 of the range in small samples (N $<$ 30) and 1/5 to 1/6 of the range in large samples (N $>$100). The formula for calculation standard deviation is as follows:

$$\sigma = \sqrt{\left\{ [\frac{1}{N}] * \sum_{i=1}^{N} (X_i - \bar{X}) \right\}} \tag{12}$$

Lets consider a couple of numerical problems to understand this concept.

1. In the *Maryland Biological Stream Survey* example , as seen earlier, the standard deviation of the data distribution will be:

   $\sigma = \sqrt{\left\{ [\frac{1}{9}] * \sum_{i=1}^{9} (X_i - \bar{X}) \right\}}$

   $= \sqrt{\left\{ [\frac{1}{9}] * \sum_{i=1}^{9} (X_i - 70) \right\}}$

   = 32.08582

2. The data shows the number of flowers per flower head of a random sample from a white clover (Trifolium repens) population.

   Data: 36, 51, 56, 62, 62, 63, 65, 69, 73, 83. Calculate the standard deviation of the flowers per flower head of this population, i.e. σ

   In this problem, to calculate the standard deviation we first need to calculate the mean of this distribution. From equation 11 we have:

   $\bar{X} = \frac{\sum X_i}{N} = \frac{(36+51+56+62+62+63+65+69+73+83)}{10}$ = 62.0

   Now we can calculate the standard deviation from equation 12 by substituting the value of mean as below:

   $= \sqrt{\left\{ [\frac{1}{10}] * \sum_{i=1}^{10} (X_i - 62) \right\}}$

   = 12.50

The above examples were a classical statistical study concerning about only one variable. Here, we observed change in one variable say (variance in flowers per

flower head, species of fishes in a stream, and etc.). However, in real world events, there exists complex relations between evidences (variables) and their corresponding outcomes(events). Thus, we have to deal with data with more than one variable.

The measure like mean, median and mode are not sufficient to analyze these in depth. There lie interesting insights to distinct appearing evidence and events which have to be explored in order to discover the hidden treasure of information. Now in order to find out the hidden, for instance, in a relation between the disease and the age of patient and/or say gender we use special statistical methods named *Correlation* and *Regression*. Thus, in the next section we present a brief study of statistical measures used for analyzing relationship between two variables.

4.3 Correlation and Linear Regression

The relationship between the variables in statistics is termed as Correlation, also known as linear correlation. The coefficient of correlation (r) is used to determine the nature of this relationship. The value of r lies between $[-1,1]$, where values between 0 to 1 imply a positively varying relation while values between -1 to 0 indicate a negative varying relation and if the value is equal to 0, then both the variables are not related or rather independent.The formula for calculating **coefficient of correlation** (r) is stated in equation 13 which is also known as the Pearson correlation coefficient.

$$r = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \tag{13}$$

A more detailed derivation of the coefficient of correlation can be found at here.[1] The correlation coefficient is used to determine the nature of the relationship within your data set. Once a correlation has been established, the actual relationship can be determined by carrying out a linear regression [3, 13].

In biology, regression and correlation are perceived in a very confused manner as pointed out by John McDonal in his work [10]. He states that in the biology domain the practice of these two methods differs primarily in its goals. Most of the researchers face challenges in the selection of which one to apply and when. He further elaborates, There are three main uses for correlation and regression in biology.

- One is to test hypotheses about cause-and-effect relationships. In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y. An example would be giving people different amounts of a drug and measuring their blood pressure. The null hypothesis would be that there was no relationship between the amount of drug and the blood pressure. If

[1] Basic statistics tutorial available online- https://controls.engin.umich.edu/wiki/index.php

the null hypothesis is rejected, the conclusion would be that the amount of drug causes changes in the blood pressure.

- The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause-and-effect relationship. In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y.
  For instance, consider a test to measure the amount of a particular protein on the surface of some cells and the pH of the cytoplasm of those cells. If the protein amount and pH are correlated, it may be that the amount of protein affects the internal pH; or the internal pH affects the amount of protein; or some other factor, such as oxygen concentration, affects both protein concentration and pH. Often, a significant correlation suggests further experiments to test for a cause and effect relationship; if protein concentration and pH were correlated, you might want to manipulate protein concentration and see what happens to pH, or manipulate pH and measure protein, or manipulate oxygen and see what happens to both.
- The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable. For example, if you were doing a protein assay you would start by constructing a standard curve. You would add the reagent to known amounts of protein (10, 20, 30 mg, etc.) and measure the absorbency. You would then find the equation for the regression line, with protein amount as the X variable and absorbence as the Y variable. Then when you measure the absorbance of a sample with an unknown amount of protein, you can rearrange the equation of the regression line to solve for X and estimate the amount of protein in the sample.

As discussed earlier, correlation examines the nature of the relation between the variables and linear regression finds a ”best fit” line that actually formulates the relation. The formula for the regression line is shown in the equation below:

$$\bar{Y} = a_0 X + a_1 \tag{14}$$

where, $\bar{Y}$ is the predicted score, a_0 and a_1 are the slope and y-intercept of the best fit line.

the formulae to find the slope and intercept of the regression line are as stated in the equation below:

$$a_0 = r\frac{\sigma_y}{\sigma_x} \tag{15}$$

$$a_1 = \bar{Y} - a_0\bar{X} \tag{16}$$

Where, r is the coefficient of correlation which is calculated using equation 13. σ_x and σ_y are the standard deviations of variables x and y respectively obtained

by substituting values in equation 12. $\bar{X}$ and $\bar{Y}$ are the mean of variable x and y respectively.

Generally, the calculations are done using statistical tools. However, to understand the concept of correlation and linear regression thoroughly let us manually examine a real world problem.

1. **The Christmas bird count**: This is a classical biology example for understanding sue of correlation and regression discussed by John McDonald in his work [10]. The data has been obtained from the Christmas bird count dataset by Audobon society[2]. There are a variety fo bird which migrate over the globe of earth in winters for breeding. The aim of this study is to find the relation between species and the distance from the equator. It is a common observation in ecology that the variety in species of the birds during winter decreases as one moves away from the equator. The data[10] is shown in the table below, which consists of the location, number of bird species and the latitude.

Location	**Latitude**	**# of species**
Bombay Hook, DE	39.217	128
Cape Henlopen, DE	38.8	137
Middletown, DE	39.467	108
Milford, DE	38.958	118
Rehoboth, DE	38.6	135
Seaford-Nanticoke, DE	38.583	94
Wilmington, DE	39.733	113
Crisfield, MD	38.033	118
Denton, MD	38.9	96
Elkton, MD	39.533	98
Lower Kent County, MD	39.133	121
Ocean City, MD	38.317	152
Salisbury, MD	38.333	108
S. Dorchester County, MD	38.367	118
Cape Charles, VA	37.2	157
Chincoteague, VA	37.967	125
Wachapreague, VA	37.667	11

Let us first plot the data on a scatter plot graph, consider figure 2 below. Here, we consider variable X as latitude and variable Y as the number of species. Now, in-order to find the correlation coefficient, we must first calculate the Mean(X) $\bar{X}$, Mean(Y) $\bar{Y}$, standard deviation of x σ_x, standard deviation of y σ_y, and various other intermediate values. In the table 1 below, we present the intermediate calculation values of the variables.

[2] The christmas bird count dataset available online http://birds.audubon.org/christmas-bird-count

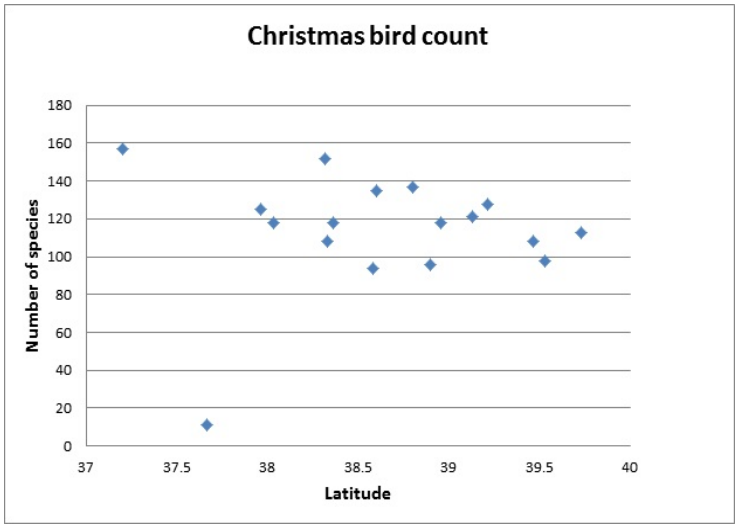

Fig. 2 The scatter plot graph of christmas bird count problem, Latitude vs Number of species.

	X	**Y**	$(X_i-\bar{X})$	$(Y_i-\bar{Y})$	$(X_i-\bar{X})^2$	$(Y_i-\bar{Y})^2$	$(X_i-\bar{X})(Y_i-\bar{Y})$
	39.216	128	0.5812	14.05882	0.3378	197.6505	8.1714
	38.799	137	0.1642	23.0588	2.6973	531.7093	3.7870
	39.4669	108	0.8312	-5.9411	0.6909	35.2975	-4.9385
	38.9579	118	0.3222	4.05882	0.1038	16.4740	1.3078
	38.6	135	-3.5764	21.0588	1.28	443.474	-0.7531
	38.5829	94	-5.2764	-19.9411	2.7841	397.650	1.0521
	39.7329	113	1.0972	-0.9411	1.2039	0.8858	-1.03269
	38.033	118	-0.6027	4.05888	0.3633	16.4740	-2.4465
	38.9	96	0.2642	-17.941	6.982	321.885	-4.7408
	39.533	98	0.8972	-15.9411	0.8050	254.1211	-14.3029
	39.133	121	0.4972	7.05882	0.2472	49.8269	3.5098
	38.317	152	-0.3187	38.0588	0.1016	1448.4740	-12.1318
	38.3323	108	-0.3027	-5.9411	9.1666	35.2975	1.7987
	38.3667	118	-0.2687	4.0588	7.2234	16.4740	-1.0908
	37.2000	157	-1.4357	43.05882	2.0614	1854.0622	-61.8223
	37.9669	125	-0.668	11.0588	0.44724	122.2975	-7.3957
	37.6670	11	-0.9687	-102.9411	0.9385	10596.8858	99.7257
SUM	656.808	1937	0	0	7.565	16338.94	8.6977
MEAN	38.6357	113.9411					

Simplifying equations 13, 12 and 15, we have:

$$a_0 = \frac{\sum[(X_i-\bar{X})(Y_i-\bar{Y})]}{\sum[(X_i-\bar{X})^2]} \tag{17}$$

Substituting values, from table 1 we find the values of the slope and the intercept, in the equations 17 and 16 as shown below:

$$a_0 = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum[(X_i - \bar{X})^2]} = 1.149$$
$$a_1 = 69.524$$

Thus, the best fit regression line from equation 14 for the data will be as:

$$a_1 = \bar{Y} - a_0 X$$
$$69.524 = \bar{Y} - 1.1496X$$
OR
$$\bar{Y} = 1.1496X + 69.524$$

Plotting the line on the scatter plot graph looks like as shown in figure below:

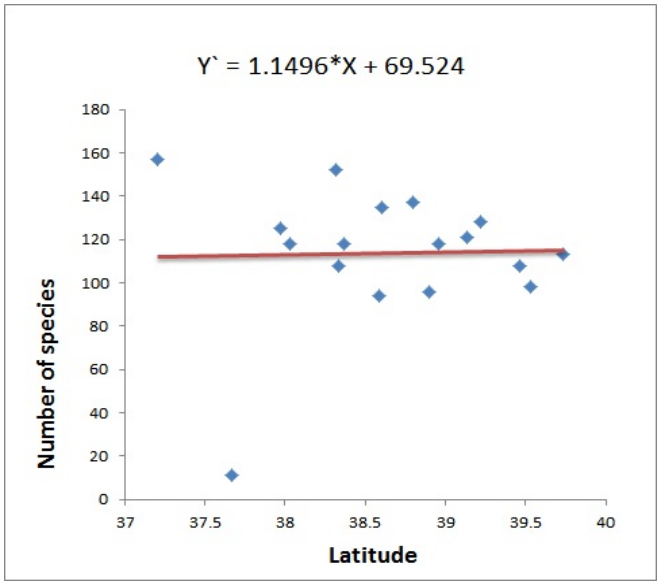

Fig. 3 The scatter plot graph of christmas bird count problem with the best fit Regression line, Latitude vs Number of species.

For further reading and practice on correlation and linear regression with more specific biology related examples suggested works are [8, 10, 12, 13].

References

1. Jaynes, E. T., 2003, Probability Theory: the Logic of Science, Cambridge University Press, see pg. 43.
2. Jeffreys, Harold (1973). Scientific Inference (3rd ed.). Cambridge University Press. p. 31. ISBN 978-0-521-18078-8.
3. Woolf, P., Burge, C., Keating, A., and Yaffe, M. (2004). Statistics and Probability Primer for Computational Biologists. Massachusetts Institute of Technology.
4. Hempel, C. G. (1945). Studies in the Logic of Confirmation (I.). Mind, 1-26.
5. Hempel, C. G. (1945). Studies in the Logic of Confirmation (II.). Mind, 97-121.
6. Raven paradox. (2014, July 21). In Wikipedia, The Free Encyclopedia. Retrieved 16:52, August 4, 2014, from http://en.wikipedia.org/w/index.php?title=Raven_paradox&oldid=617899290
7. Darling D., The Ravens paradox published online at http://www.daviddarling.info/encyclopedia/R/raven_paradox.html
8. Hanon B., and Larget B., Lecture notes on Statistics 571 available online at www.stat.wisc.edu/ st571-1/
9. Stansfield, W. D., and Carlton, M. A. (2004). Bayesian Statistics for Biological Data: Pedigree Analysis. The American Biology Teacher, 66(3), 177-182.
10. McDonald, J. H. (2009). Handbook of biological statistics (Vol. 2, pp. 173-181). Baltimore, MD: Sparky House Publishing.
11. Lane, D. (2011). Online Statistics Education. In International Encyclopedia of Statistical Science (pp. 1018-1020). Springer Berlin Heidelberg.
12. StatSoft, I. (2007). Electronic statistics textbook. StatSoft, Tulsa, OK.
13. Krickeberg, K., Pham, T. M. H., and Pham, V. T. (2012). Epidemiology. Springer.